International Journal of Youth Economy

Int. J. Youth Eco. 1, No. 1, 75-104 (2017)
DOI. http://dx.doi.org/10.18576/ijye/010106

What Controls GDP: Population, Area & Rail Lines?

*Shady M. Qubaty**

Yale University '20

Received: 3 Feb. 2017, Revised: 2 Mar. 2017, Accepted: 4 Mar. 2017.
Published online: 1 May 2017.

Abstracts

This paper investigates the relationship between GDP and the population of a country, the total length of rail lines and the total area. It makes intuitive sense that these variables should all have a relationship with the total gross output of a country (GDP). However, this paper will examine the extent of this association. After having conducted a pilot study of 15 randomly selected countries, it became clear that there exists some form of positive relationship. Through the use of various statistical methods, it turns out that there is a stronger positive correlation between the population and total rail lines of a country and its GDP than the total area.

Keywords: Economics, GDP, Area, Population, Rail, Correlation, Causation, Association.

Investigating Hypothesis

Gross Domestic Product (GDP), is arguably the most important of all economic statistics as it attempts to capture the state of the economy in one number. It is a measure of the overall economic output within a country's borders over a particular time, typically a year. Likewise, it is the sum of all goods and services produced in the economy, including the service sector, manufacturing, construction, energy, agriculture and government. However, the GDP is calculated without making deduction for the depreciation of fabricated assets or for depletion and degradation of natural resources.

This investigation will examine the relationship between the GDP of different countries and three vital factors. These factors which intuitively seem to directly affect the GDP are the: population, area and the total length of the rail lines of a country. Whenever the paper references GDP, it implies GDP (PPP) which is the Gross Domestic at Product Purchasing Power Parity. The advantage of using GDP (PPP) in comparison to the real GDP is that it removes exchange rate fluctuations.

Those are the hypotheses, which will be tested throughout this investigation:

a) The higher the population of a country, the higher its GDP.

*Corresponding author e-mail: shady.qubaty@yale.edu

b) The bigger the area of a country, the higher its GDP.

c) The higher the GDP of a country, the longer the length of its total rail lines.

Data collection and sampling methods

In reference to the hypotheses, correlation is a measure of the strength of the linear association between two variables; it can either be a strong linear correlation (where the points lie close to the regression line) or a weak linear correlation (where the points lie scattered away from the regression line). Notably, there are two types of correlation, negative correlation, which is said to exist when one covariate increases as the other covariate decreases, and positive correlation which is when one factor increases as the other factor increases. Both negative and positive correlation can either have a weak or a strong linear correlation according to the position of the points relative to the line of best fit.

The primary concern of using secondary data in any investigation is that it may not come from a trusted and reliable source. One can never be sure about the accuracy of the secondary data provided as it may often be out of date. Furthermore, any data (primary or secondary) may be biased as it may not represent the whole population. For instance, the data will be limited to certain countries only because the statistics of some countries will not be available, thus it does not represent the world as a whole. Also, the data may contain outliers and anomalies, which this investigation will try to exclude from the data to make the sample more representative and unaffected by those extremities.

Apart from that, secondary data can be put under two categories; either quantitative data or qualitative data. The data provided in the World Bank database is quantitative as it is all made up of numerical measurements. Furthermore, quantitative data can be divided into two types of data; either discrete data or continuous data. Discrete data can only take a particular value on a numerical scale (integer) and is represented by something that is countable, for instance the number of countries in the world (you cannot say that there are 190.5 countries in the world, it’s either 190 or 191). Continuous data is what this investigation uses as it can take any value on a continuous numerical scale, even if you cannot measure an exact value for it, for example the GDP of a country ($16,193,423.87).

For studying both these hypotheses not only will one resort to the World Bank database, but one needs to find a suitable sampling method as it is not always the best option to take the whole population because that will make it harder to analyze the data and to identify anomalies. Many sampling methods will be available for me, for example: systemic sampling, purposive sampling, opportunity sampling, cluster sampling and random sampling etc. The author decided to use Random sampling because it is the most appropriate method to use for the secondary data and also because it reduces bias as the method is completely random. Random sampling is made up of two types: stratified sampling and simple random sampling. Stratified sampling is used to make sure that the size of the sample taken is in proportion to the relative size of the stratum from which the data is taken. The second type is simple random sampling which is when each sample of size has the same probability of being selected. To get a random set of numbers, one will begin by finding all the countries which have the figures for the data needed and will give each country a number on the Excel Spreadsheet. After that, this investigation will use a random number generator using the website www.random.org to generate 60 random numbers which will determine the countries used for the investigation. If the same number is repeated twice, then one should generate another number.

Diagrams and Calculations

Throughout the investigation, a wide variety of statistical techniques will be used through Microsoft Excel. By carrying out those relevant statistical techniques, one will be able to observe the association between the variables in the hypotheses.

To process the raw data, sort it and to spot the patterns in the data, one will start by making frequency distribution tables for each of the variables. It will make the data easier to interpret and it is beneficial because continuous data will be able to be sorted in the table in a manageable manner. The frequency tables will allow the author to calculate averages for the data. It will give the author the chance to find the mean, median interval, modal interval, range and the cumulative frequency. The mean of the data will tell the author approximately the average of the data. The median interval will tell the author where the middle value of the data lies, the modal interval will tell the author where most of the data will lie and the range will determine how spread the data will be. The cumulative frequency of the value of a variable is the total number of observations that are less or equal to that value and that will help the author plot a cumulative frequency diagram for each of the variables.

Using the frequency table, one will plot a bar chart for each of the three variables which will give the author the opportunity to visualize the bivariate data (pairs of related variables). Unlike the frequency tables, values can be read from the scale in a bar chart which will make it easier for the readers to visualize the data.

The author will add an extra column for the frequency distribution tables and will call it the “frequency density”. That will be obtained by dividing the frequency by the class width. By finding the frequency densities, the author will be able to construct a histogram for each of the variables which will be a graphical display of the tabulated frequencies. A histogram shows how the data are distributed across the class intervals. It is similar to a bar chart but because the data is continuous, there are no gaps between the bars. It gives you an idea on what proportion of classes fall into each of several or many specified categories. The histogram differs from a bar chart in that it is the area of the bar that denotes the value, not the height. Hence, it will give the author the chance to find the probability of a certain interval. The histogram will also give the author the opportunity to find out how the data is distributed across the class intervals. It will tell the author if the distribution is symmetrical, has a positively skewed or negatively skewed.

Along with the bar chart, using the cumulative frequency section of the frequency tables, the author will be plotting cumulative frequency diagrams for each of the variables which will give the author the opportunity to estimate or predict additional values. The author will draw the cumulative frequency diagram using the computerized software AUTOGRAPH to make the results more reliable than just drawing the graph by hand. A cumulative frequency diagram is drawn by plotting the cumulative frequencies against their corresponding upper class boundaries. It will make it easier for the reader to apprehend the data. Using the cumulative frequency diagrams, the author will be able to find the minimum value, maximum value, lower quartile (LQ), median, upper quartile (UQ), inter-quartile range (IQR), percentiles, inter-percentile range (IPR), deciles and interdecile range (IDR). The lower quartile is the value such that one quarter (25%) of the values are less than or equal to it. The upper quartile is the value such that three quarters (75%) of the values are less than or equal to it. By subtracting the LQ from the UQ one will be able to obtain the IQR which will identify how spread the middle 50% of the data is. The good thing about the IQR is that it's not affected by extreme values unlike the range. The median is the value such that half (50%) of the values are less than or equal to it. Unlike the median interval that is found using the frequency diagrams, the median found here will tell the author the exact middle value which is much more accurate. Percentiles are used to divide data into 100 groups and deciles are used to divide data into 10 groups. While quartiles split data up in quarters (in jumps of 25%) percentiles and deciles are more general methods which split it into 100 groups and 10 groups (in jumps of 1% and 10%). The IPR will determine between what percentiles the data are most concentrated. The cumulative frequency diagram will also make the author able to determine any anomalies because one will be able to see if the piece of datum follows the trend or not.

Using the values one obtains from the cumulative frequency diagram, the author will be able to plot a box plot diagram for each of the variables to help illustrate the data obtained from the cumulative frequency diagram. The author will be using the software AUTOGRAPH to construct the box plots as it will be more accurate than drawing it by hand. Box plots represent important features of the data; the maximum and minimum values, the median, and the upper and lower quartiles. These different factors will give the author the opportunity to compare the data. The box plots will show the author whether the distribution of the data is symmetrical, positively skewed or negatively skewed. If the median is closer to the UQ than it is to the LQ the data is negatively skewed. If the median is closer to the LQ than it is to the UQ the data is positively skewed and if the median is equidistant from both the upper and lower quartiles the data is symmetrical. Using the IQR obtained from the cumulative frequency diagram, one will be able to accurately determine if the data contains any outliers.

To see if there is any connection between all of the 60 samples, one will draw a scatter diagram for each hypotheses stated. A scatter diagram is a tool for analyzing relationships between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis. The pattern of their intersecting points can graphically show relationship patterns. Most often, a scatter diagram is used to prove or disprove cause-and-effect relationships which, and that is the main reason it will be used. The scatter diagrams will be drawn just like the ones drew for the pilot study but this on a sample of 60 instead of only 15. The scatter diagram will show the author the association between the variables in the hypotheses and this will be the most important stage of this investigation as one will be able to observe if the hypotheses stated are correlated. After one plots all the points on the scatter diagram, a line of best fit will determine the general trend of the data and it will allow the author to predict results by extrapolating the data. Consequently, the author will find the equation of the line of best fit, to determine the general trend of the hypotheses in a numerical manner.

To view the correlation of the variables in the hypothesis in a quantitative and more precise manner, the Spearman's Rank Correlation Coefficient will be used, just like the one used in pilot study, but this on the full sample of 60. It will give the author a value between -1 and 1, which will inform the author on the strength of correlation between the variables. The closer the number obtained is to 1 the more the positive correlation, the closer the number is to -1 the more the negative correlation and if the number is close to 0, that implies that the variables are not correlated.

Population

Total Population(100,000)	Frequency	Midpoint(x)	fx	Cumulative frequency (c.f)
0<x≤20	29	10	290	29
20<x≤40	10	30	300	39
40<x≤60	6	50	300	45
60<x≤80	6	70	420	51
80<x≤100	3	90	270	54
100<x≤120	1	110	110	55
120<x≤140	1	130	130	56
Total:	**∑f=56**		**∑fx=1820**	

Outliers have been excluded where x>140

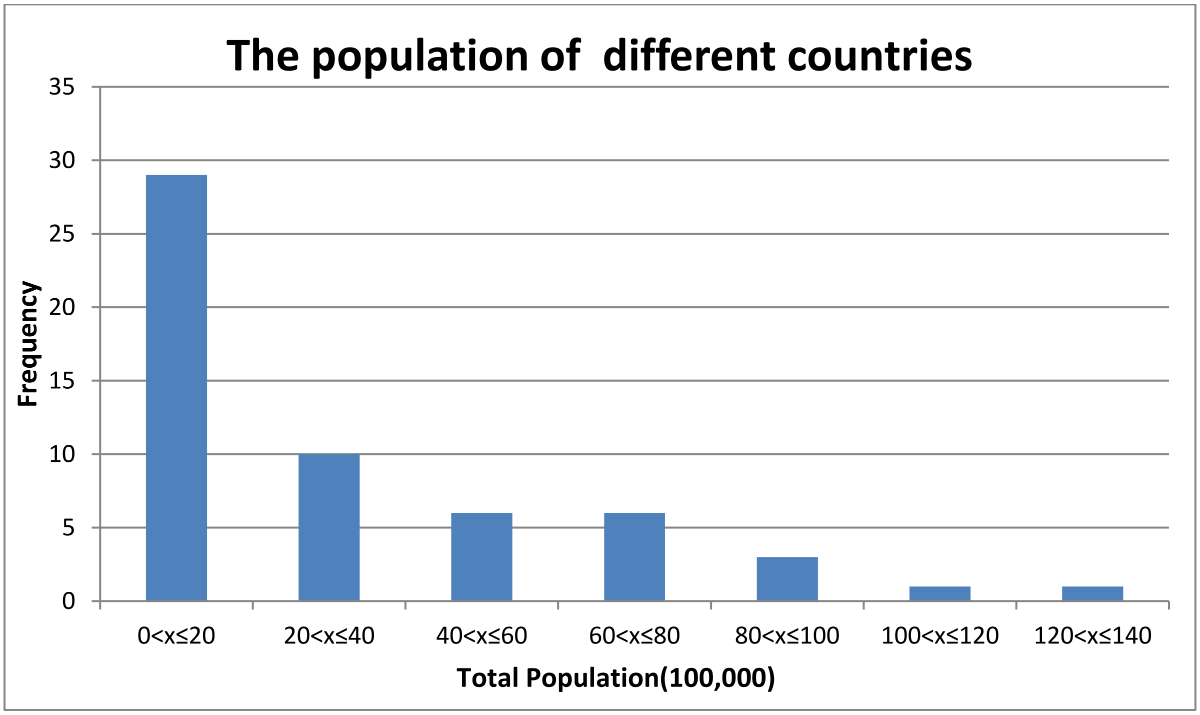
$Mean = \frac{1820}{56}$ =32.5 hundred thousand people= **3,250,000 people**

Median interval =$\left(\frac{n}{2}\right)th\ position$, n= 56 ∴$\frac{56}{2} = 28th\ position$ = **0<x≤2,000,000 people**

Modal interval = **0<x≤2,000,000 people**

Range = highest value – lowest value ∴ 14,000,000 - 0= **14,000,000 people**

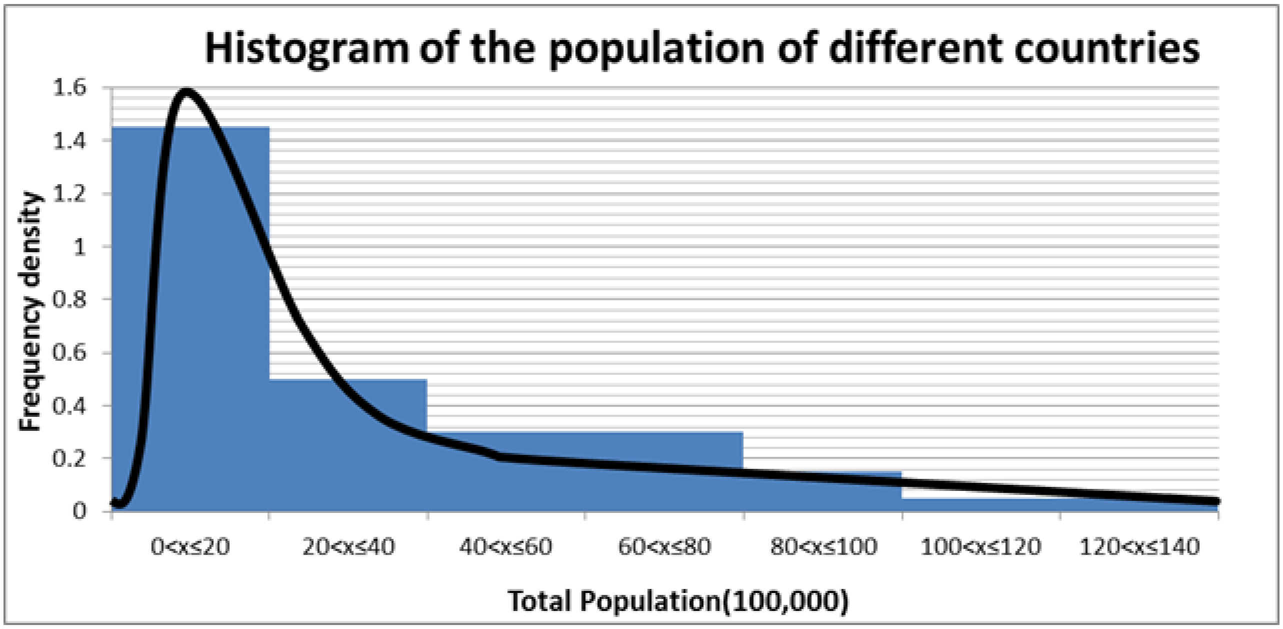
On average (excluding anomalies) a country has about 3,250,000 people. Most of the countries in the sample have a population between 0 and 2,000,000 people.



From the bar chart above, we can see that the majority of the countries in the investigation have a population between 0 and 20,000,000 people. From the bar chart we can also observe that the frequency of the countries decreases as the total population size increases. The bar chart gives us the chance to compare different intervals visually. For instance, from this chart we can see that in the investigation, only very few countries had a population more than 100,000,000 people.

Total Population(100,000)	Frequency (f)	Class width (cw)	Frequency density (fd)
0<x≤20	29	20	1.45
20<x≤40	10	20	0.5
40<x≤60	6	20	0.3
60<x≤80	6	20	0.3
80<x≤100	3	20	0.15
100<x≤120	1	20	0.05
120<x≤140	1	20	0.05

Using the frequency table above which includes the frequency densities for each interval a histogram for the population of the countries included in the investigation could be created. If you look at the frequency column of table you will realize that they do not add up to 60 and that is because some outliers were excluded to make the results more representative and reliable:



From the histogram we can see that there is a positive skew as most of the data values are at the lower end. This tells us that most of the countries in the investigation have a total population which is low (excluding outliers).

From the histogram, we can identify the probability of a certain interval. To find out what percentage of the whole area the first interval occupies, we will find the area of that interval and divide it by the total area of all intervals than multiply it by 100 to get it as a percentage.

Area of the first interval= 1.45 x 20,000,000= 29,000,000 people

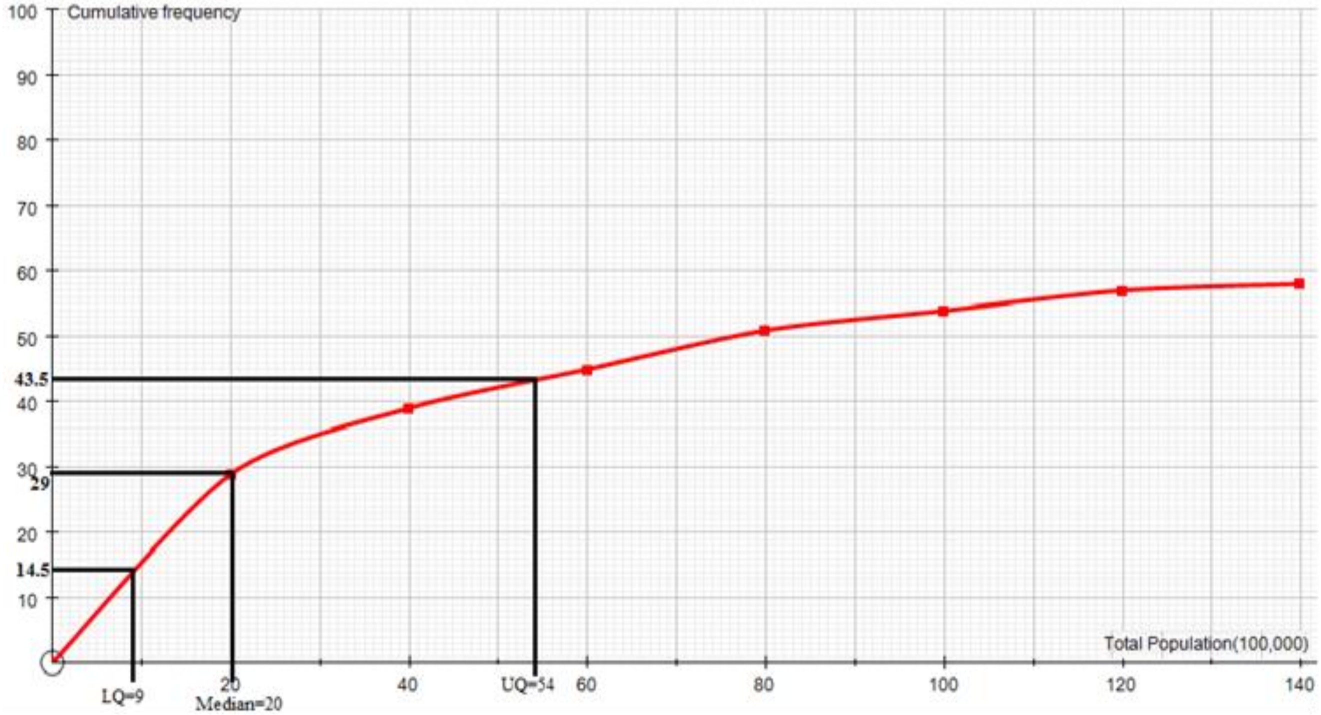
The area of all the intervals= (1.45+0.5+0.3+0.3+0.15+0.05+0.05)× 20,000,000

= 2.8 x 20,000,000= 56,000,000 people ∴ probability= $\frac{29{,}000{,}000}{56{,}000{,}000} \times 100 = 51.8\%$

From finding the probability of the first interval, more than half (51.8%) of the countries in the investigation (excluding outliers) have a population between 0 and 2,000,000 people. This supports the distribution of the data that identified previously because most of the data values are at the lower ends.

Total Population(100,000)	**Frequency**	**Cumulative frequency (c.f)**
0<x≤20	29	29
20<x≤40	10	39
40<x≤60	6	45
60<x≤80	6	51
80<x≤100	3	54
100<x≤120	1	55
120<x≤140	1	56

Now, using the following table, we can plot a cumulative frequency diagram:



From this cumulative frequency graph we can calculate many things:

LQ=$\left(\frac{n}{4}\right)$th value= $\frac{58}{4}$ =14.5th value= 900,000 people

Median=$\left(\frac{n}{2}\right)$th value= $\frac{58}{2}$ =29th value= 2,000,000 people

UQ=$\left(\frac{3n}{4}\right)$th value= $\frac{174}{4}$ =43.5th value= 5,400,000 people

IQR= UQ-IQ= 5,400,000-2,000,000=3,400,000 people

Maximum Value = 14,000,000 people

Minimum Value =50,000 people

Above when calculating the median interval it was **$0<x\leq2{,}000{,}000$ people** and now when calculating the actual median it is 2,000,000 people which shows that the median obtained from the cumulative frequency graph is correct as it falls between the median interval previously calculated.

While quartiles divide the data into 4 groups, deciles are more general and divide the data into 10 groups. Percentiles are the most general method as they divide the data into 100 groups. One type of percentile calculation is deciles. Deciles are a percentile taken in tens. The first decile is equivalent to the 10th percentile and the second decile is the 20th percentile, and so forth.

The general formula for deciles is: dth $decile = \left(\frac{(D)n}{100}\right)$, where "n" is the total number of observations, "d" is a number between 1 to 10 and "D" is a number in its tens like 10, 20,30,40,50,60,70,80,90 and 100

For instance to find the 2nd decile of the data:

2nd decile=$\left(\frac{20(58)}{100}\right)$= 11.6th value=700,000 people.

We can also calculate the IDR from (interdecile range) from the cumulative frequency graph. The interdecile range is the difference between the first and the ninth deciles. The interdecile range is a measure of statistical dispersion of the values in a set of data, similar to the range and the interquartile range.

It is the 10th to 90th IPR. So, the IDR= P_{90}-P_{10}

P_{90}=$\left\{\frac{90(58)}{100}\right\}$th and P_{10}=$\left\{\frac{10(58)}{100}\right\}$th ∴ P_{90}=52.2nd value and P_{10} = 5.8th value.

P_{90}=9,000,000 and P_{70}=400,000 ∴ IDR= 9,000,000-400,000=**8,600,000 people.**

- One can use the data obtained from the population cumulative frequency diagram to plot a box and plot diagram.

Measures of spread	**Cumulative frequency population (100,000)**
Minimum value	0.5
LQ	9
Median	20
UQ	54
IQR	34
Maximum value	140

Before constructing the box and plot diagram, one has to first identify the outliers

To see if there are any outliers:

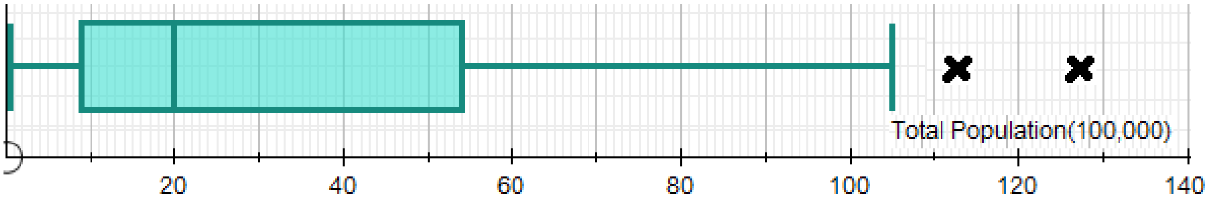
x= IQR × 1.5 = 34×1.5=51

- Highest value not an outlier (H)= upper quartile + x = 54+51=105

Any value greater than 105(>105) is an outlier. Therefore, one will exclude them from the data to make the results more accurate and representative. Listed below are the two outliers that will be excluded from the data.

Country	Total Population (100,000)
Mexico	113.4
Japan	127.5

These are the outliers in the data and will be excluded. They will be plotted in the box plot diagram as crosses.



Median =29 and Median-LQ=13, subsequently we can see that the median is closer to the LQ then to the UQ and that means that the data is positively skewed. That agrees with the skewness found in the histogram for the "population" variable.

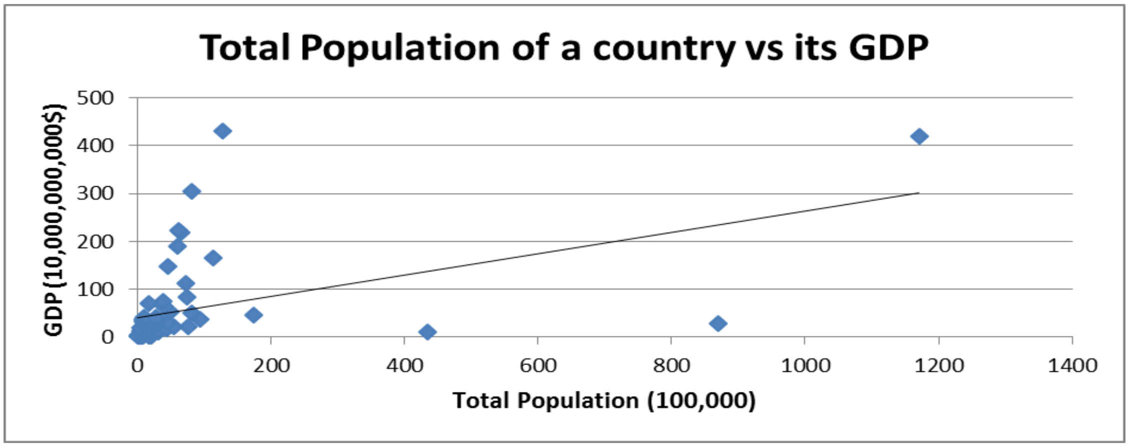
The # column shows the random numbers of each country that were selected randomly from the whole sample.

#	Country	Total Population (100,000)	GDP (10,000,000,000$)	Rank Population	Rank GDP	d	d^2
56	Moldova	3.7	1.11	8	1	7	49
57	Mongolia	2.8	1.12	6	2	4	16
48	Kyrgyz Republic	5.4	1.23	11	3	8	64
11	Benin	8.8	1.4	18	4	14	196
81	Tajikistan	6.9	1.49	14	5	9	81
4	Armenia	3.1	1.69	7	6	1	1
24	Congo, Rep.	4	1.73	9	7	2	4
17	Burkina Faso	16.5	2.18	28	8	20	400
34	Georgia	4.5	2.26	10	9	1	1
33	Gabon	1.5	2.29	3	10	-7	49
52	Macedonia, FYR	2.1	2.3	4	11	-7	49

30	Estonia	1.3	2.76	2	12	-10	100
14	Botswana	20.1	2.79	30	13	17	289
49	Latvia	2.2	3.66	5	14	-9	81
51	Luxembourg	0.5	4.39	1	15	-14	196
12	Bolivia	9.9	4.81	22	16	6	36
70	Serbia	7.3	8.23	15	17	-2	4
7	Azerbaijan	9	8.99	19	18	1	1
68	Sudan	435.5	9.82	58	19	39	1521
82	Tunisia	10.5	10.1	23	20	3	9
75	Sri Lanka	20.9	10.59	32	21	11	121
80	Syria	20.4	10.81	31	22	9	81
40	Iraq	32	11.41	38	23	15	225
71	Slovak Republic	5.4	12.73	12	24	-12	144
9	Belarus	9.5	13.22	21	25	-4	16
58	Morocco	31.9	15.31	37	26	11	121
41	Ireland	44.8	18.46	41	27	14	196
31	Finland	5.4	19.66	13	28	-15	225
46	Kazakhstan	16.3	19.86	27	29	-2	4
42	Israel	76.2	21.77	51	30	21	441
28	Denmark	55.4	21.89	45	31	14	196
27	Czech Republic	10.5	26.61	24	32	-8	64
63	Peru	29.1	27.73	36	33	3	9
90	Vietnam	869.4	27.86	59	34	25	625
67	Romania	21.4	30.63	33	35	-2	4
85	Ukraine	45.9	30.83	42	36	6	36
36	Greece	11.3	31.47	26	37	-11	121
6	Austria	8.4	33.54	17	38	-21	441
89	Venezuela	28.8	35.27	35	39	-4	16
79	Switzerland	7.8	36.45	16	40	-24	576
78	Sweden	9.4	36.61	20	41	-21	441
64	Philippines	93.3	37	54	42	12	144
10	Belgium	10.9	40.91	25	43	-18	324
53	Malaysia	28.4	41.84	34	44	-10	100
62	Pakistan	173.6	46.66	57	45	12	144
29	Egypt	81.1	50.13	52	46	6	36
73	South Africa	50	52.84	44	47	-3	9

3	Argentina	40.4	64.71	40	48	-8	64
60	Netherlands	16.6	70.19	29	49	-20	400
65	Poland	38.2	75.55	39	50	-11	121
39	Iran	74	84.62	50	51	-1	1
83	Turkey	72.8	111.46	49	52	-3	9
74	Spain	46.1	147.78	43	53	-10	100
55	Mexico	113.4	164.44	55	54	1	1
43	Italy	60.5	190.86	46	55	-9	81
32	France	64.9	219.41	48	56	-8	64
86	United Kingdom	62.2	223.39	47	57	-10	100
35	Germany	81.7	304.42	53	58	-5	25
38	India	1170.9	419.49	60	59	1	1
44	Japan	127.5	430.18	56	60	-4	16
Σn= 60						***Σd²* =8990**	

Using the table, one could construct a scatter diagram:



From the scatter graph above we can see that although there are a few anomalies there is still a positive correlation between the two variables (population and GDP). To mathematically prove that there is a positive correlation, one can find the gradient of the line of best fit using the following formula:

The two co-ordinates that will be used to find the gradient of the line of best fit are (300,100) and 750,200).

$$\frac{750-300}{200-100} = \frac{450}{100} = 4.5x$$

The gradient is positive and that tells us that there is a positive correlation between the two variables of the hypothesis. But to find exactly the strength in the correlation between those two variables, one needs to calculate S.R.C.C:

S.R.C.C = $1-\frac{6(8990)}{60(60^2-1)} = 1-\frac{53940}{60(3599)} = 1-\left(\frac{53940}{215940}\right)$

= 1-0.25= **0.75** ∴ **S.R.C.C= 0.75**

S.R.C.C shows that the variables are strongly positively correlated. From the scatter diagram, one was able to say that there is a positive correlation but now we can say that there is a strong positive correlation. In comparison to the S.R.C.C obtained from the pilot study on the same variables, the S.R.C.C here is larger (by 0.125). That is because the sample here is bigger and therefore the results are more accurate and representative.

Area

Constructing a frequency distribution table for the "area of a country":

Area (10,000 sq. km)	**Frequency (f)**	**Midpoint(x)**	**fx**	**Cumulative frequency (c.f)**
0<x≤25	28	12.5	350	28
25<x≤50	15	37.5	562.5	43
50<x≤75	3	62.5	187.5	46
75<x≤100	4	87.5	350	50
100<x≤125	2	112.5	225	52
125<x≤150	1	137.5	137.5	53
150<x≤175	2	162.5	325	55
175<x≤200	1	187.5	187.5	56
Total:	**∑f**= 56		**∑fx**=2325	

$Mean = \frac{2325}{56}$ =41.5 ten thousand sq. km = **415,000sq km**

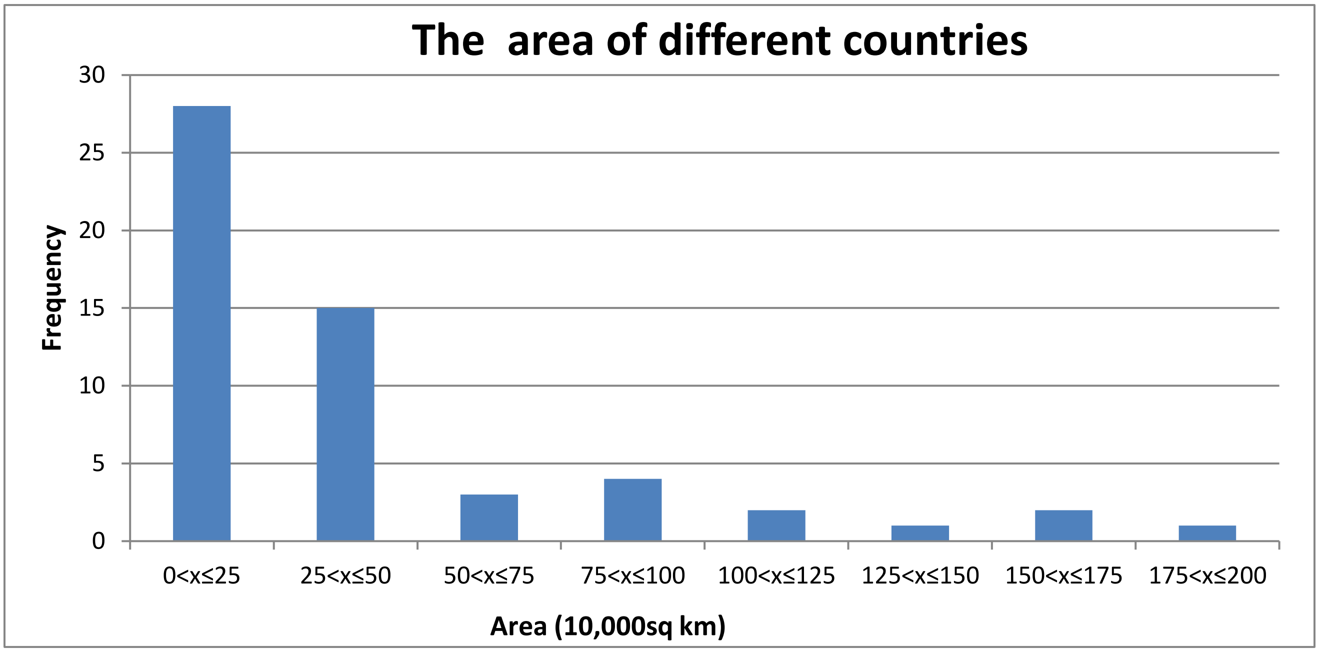
Median interval =$\left(\frac{n}{2}\right)$ $th\ position$, n= 56 ∴$\frac{56}{2}$ = 28$th\ position$ = **0<x≤250,000sq km**

Modal interval = **0<x≤250,000 sq. km**

Range = highest value – lowest value ∴ 200,000-0= **200,000sq km**

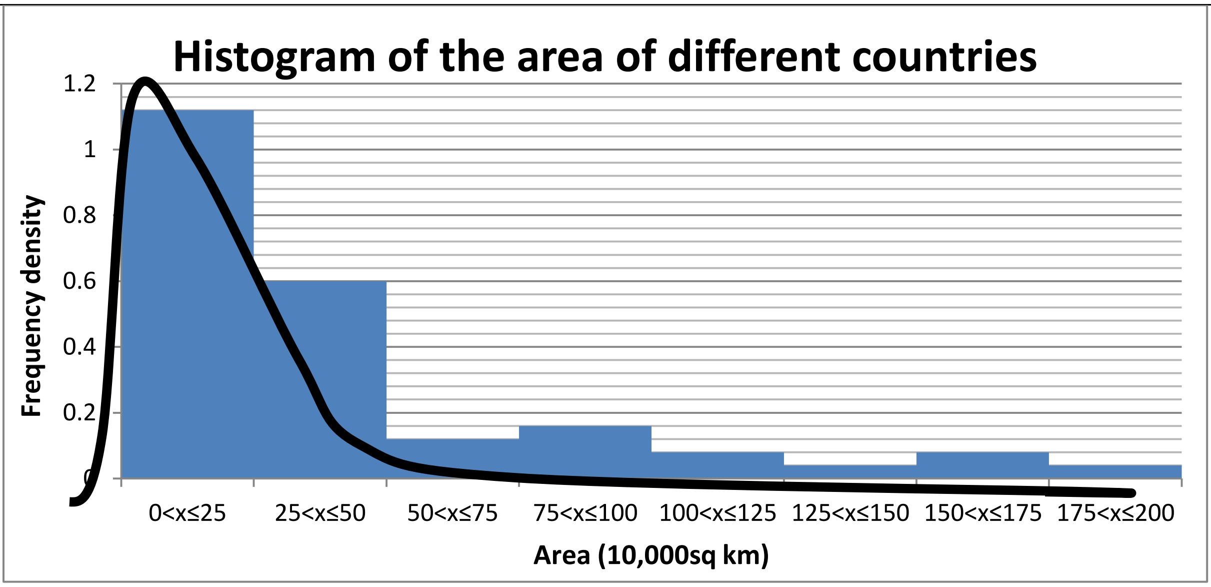
The averages found tell us really important features about the continuous data. One can also see that on average (excluding outliers); a countries area is about 200,000sq km. Most countries in the sample have an area between 0 and 250,000sq km.

- One can draw a bar chart using the data from the frequency distribution tables. It will visually show the the frequencies of the area intervals of the countries included in the investigation, excluding outliers:



From the bar chart we can see that the majority of the countries included in the investigation have an area between 0 and 250,000 sq. km and the minority have an area between 1,250,000 and 1,500,000sq. km. The good thing about a bar chart is that it makes it easier for the reader to visualize your data and it is very simple to understand what it is showing, unlike a frequency distribution diagram which many people find hard to interpret the data shown on it.

Area (10,000 km)	Frequency (f)	Class width (cw)	Frequency density(fd)
0<x≤25	28	25	1.12
25<x≤50	15	25	0.6
50<x≤75	3	25	0.12
75<x≤100	4	25	0.16
100<x≤125	2	25	0.08
125<x≤150	1	25	0.04
150<x≤175	2	25	0.08
175<x≤200	1	25	0.04



From the histogram above we can see that this distribution is positively sewed which means that most of the values are at the lower end of the intervals. It shows that most of the countries in the investigation have a small area (excluding anomalies).

To find the probability of the first interval which is the largest, one needs to find the area of that interval and then divide it by the total area of all the intervals before multiplying it by 100 to get the probability of that interval as a percentage.

Area of the first interval= 1.12 x 250,000 = 280,000 sq. km

The area of all the intervals= (1.12+0.6+0.12+0.16+0.08+0.04+0.08+0.04) × 250,000

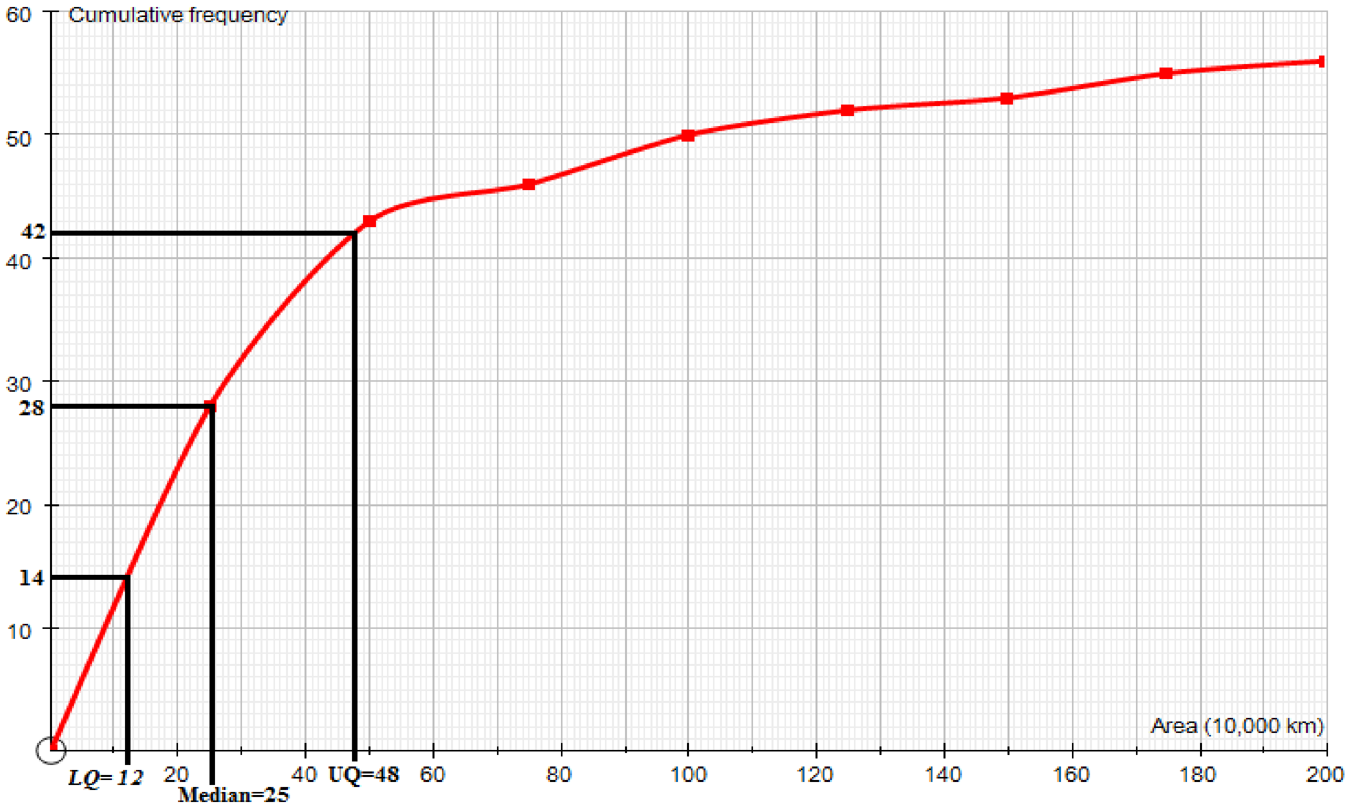
= 2.24 x 250,000= 560,000 sq. km ∴ probability= $\frac{280{,}000}{560{,}000} \times 100 = 50\%$

By finding the probability of the first interval, one can say that 50% of the countries in the investigation (excluding outliers) have an area between 0 and 250,000 sq. km. That's the great thing about histograms; it gives you the opportunity to find the probability of a certain interval.

- We will construct the table for the main factor in the area hypothesis. One can use this table to plot the cumulative frequency diagram. This table will include the intervals of the variable, the frequencies of each variable and the cumulative frequency (c.f).

Area (10,000 km)	**Frequency**	**Cumulative frequency (c.f)**
0<x≤**25**	28	**28**
25<x≤**50**	15	**43**
50<x≤**75**	3	**46**
75<x≤**100**	4	**50**
100<x≤**125**	2	**52**
125<x≤**150**	1	**53**
150<x≤**175**	2	**55**
175<x≤**200**	1	**56**

The total frequency does not add to 60 because some outliers have been removed. Now by



using the upper value of the intervals and the c.f, one would be able to draw a cumulative frequency diagram:

From the cumulative frequency diagram, we can calculate many things:

LQ=$\left(\frac{n}{4}\right)^{th}$ value= $\frac{56}{4}$=14th value= 120,000sq km

Median=$\left(\frac{n}{2}\right)^{th}$ value= $\frac{56}{2}$=28th value= 250,000sq km

UQ=$\left(\frac{3n}{4}\right)^{th}$ value= $\frac{168}{4}$=42nd value= 480,000sq km

IQR= UQ-IQ= 480,000-12,000=36,000

Maximum Value = 200,000sq km

Minimum Value =3000sq km

Above when calculating the frequency distribution, the median interval for the area was $0<x\leq 250{,}000$sq km and here we can see that the exact median is 250,000sq km. That shows that the median obtained here is correct as it lies within the median interval previously found.

While the quartiles split the data up in quarters, there is a more general method which splits it into 100 groups called percentiles. Percentiles divide data into 100 equal parts. They are commonly used for providing a relative standing of an event or person in a population. You can also calculate percentiles for grouped data.
The general formula for percentiles is: pth $percentile = \left(\frac{Pn}{100}\right)$, where n is the total number of observations. For the area, the total number of observations is 56 because 4 entries have been removed as outliers.

For example in this cumulative frequency diagram, to find the 7th percentile:

7th percentile=$\frac{7(56)}{100}$ = 3.9th value =30,000sq km.

We can also find the IPR (interpercentile range) which is the difference between two percentiles. The inter-percentile range is a stable measure of spread (unless one of the percentiles is the minimum or maximum), meaning that the value is quickly obtained for relatively few iterations of a model. It also has the great advantage of having a consistent interpretation between distributions. The IQR (interquartile range) is the IPR between the 25th and 75th percentiles.

So for instance to find the IPR between the 70th and 90th percentile that's what we do:

70th to 90th IPR= P_{90}-P_{70}

P_{90}= $\left\{\frac{90(56)}{100}\right\}^{th}$ and P_{70}=$\left\{\frac{70(56)}{100}\right\}^{th}$ ∴ P_{90}=50th value and P_{70} = 39th value.

P_{90}=1,000,000 and P_{70}=420,000 ∴ IPR= 1,000,000-420,000=580,000sq km.

- We will start by constructing a box and plot diagram for the first variable which is the "area" of a country. One can use the information obtained from cumulative frequency diagram to make this box and plot diagram. To construct it, one needs to know, the lowest value, LQ, median, UQ and the highest value:

Measures of spread	**Cumulative frequency area (10,000sq km)**
Minimum value	0.3
LQ	12
Median	25
UQ	48
IQR	36
Maximum value	200

To see if there are any outliers:

x= IQR × 1.5 = 36×1.5=54

- Highest value not an outlier (H)= upper quartile + x = 48+54=102

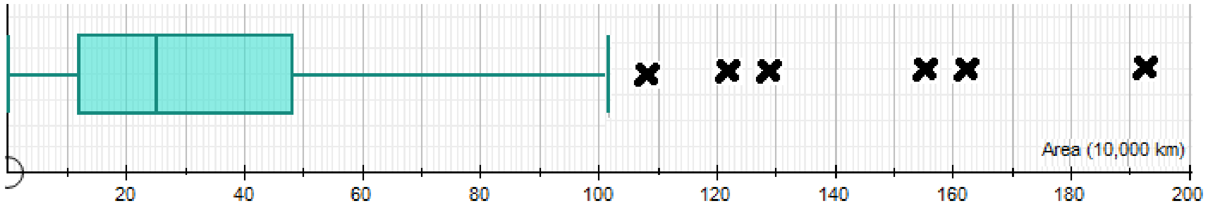
Any value greater than 102(>102) is an outlier. Therefore, those entries will be excluded them from the data to make the results more accurate and representative:

-

Country	Land area (10,000sq. km)
Bolivia	108.3
South Africa	121.4
Peru	128
Mongolia	155.4
Iran	162.9
Mexico	194.4

These are the outliers in the data which will be excluded

- Using the measures of spread obtained, one can plot a box plot which will display all the data. On the box plot, we can also mark the outliers as crosses. The maximum value of the box plot will be the upper outlier because all of the data exceeding that point are considered as outliers. So for the area, the maximum value will be 102 and not 200:



UQ - Median = 23 and Median – LQ= 13, therefore the median is closer to the LQ than it is to the UQ and that shows that the data is positively skewed.

The # column shows the random numbers of each country that were selected randomly from the whole sample.

#	Country	Land area (10,000sq. km)	GDP (10,000,000,000$)	Rank Area	Rank GDP	d	d²
56	Moldova	3.3	1.11	6	1	5	25
57	Mongolia	155.4	1.12	54	2	52	2704
48	Kyrgyz Republic	19.2	1.23	25	3	22	484
11	Benin	11.1	1.4	20	4	16	256
81	Tajikistan	14	1.49	22	5	17	289
4	Armenia	2.8	1.69	4	6	-2	4
24	Congo, Rep.	34.2	1.73	37	7	30	900
17	Burkina Faso	27.4	2.18	30	8	22	484
34	Georgia	6.9	2.26	14	9	5	25

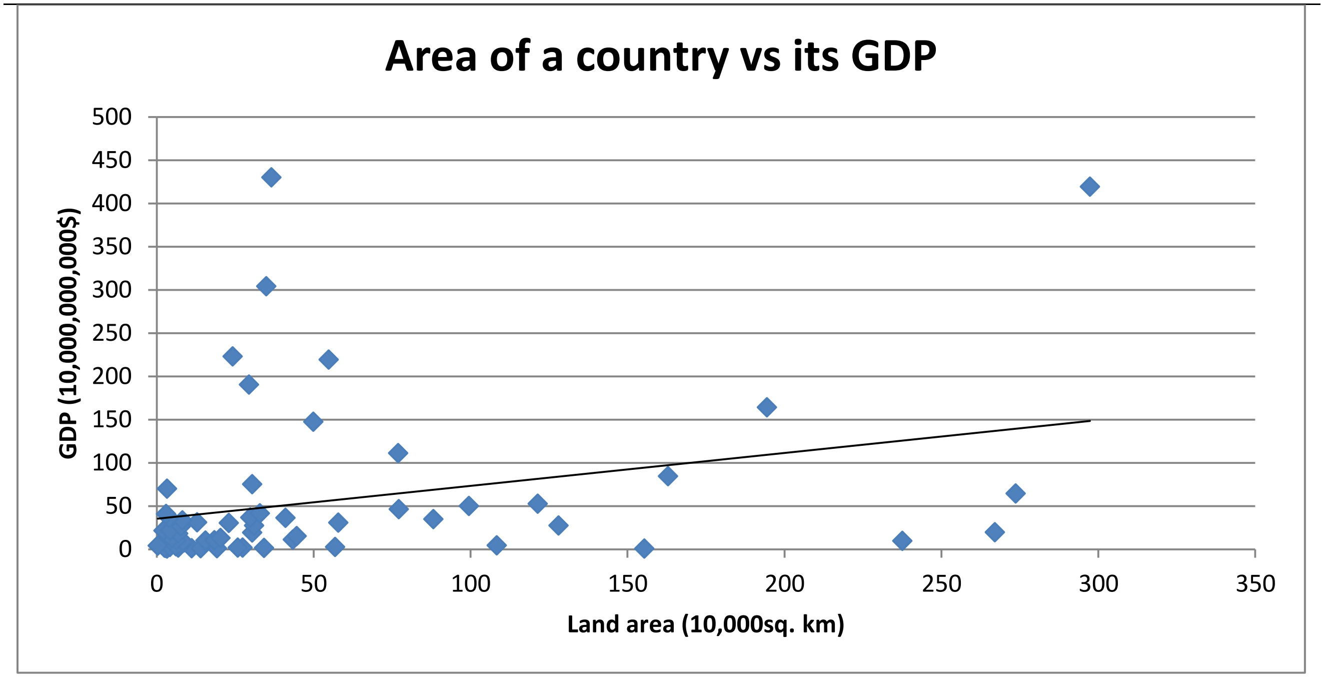
33	Gabon	25.8	2.29	29	10	19	361
52	Macedonia, FYR	2.5	2.3	3	11	-8	64
30	Estonia	4.2	2.76	9	12	-3	9
14	Botswana	56.8	2.79	45	13	32	1024
49	Latvia	6.2	3.66	12	14	-2	4
51	Luxembourg	0.3	4.39	1	15	-14	196
12	Bolivia	108.3	4.81	51	16	35	1225
70	Serbia	8.7	8.23	19	17	2	4
7	Azerbaijan	8.3	8.99	18	18	0	0
68	Sudan	237.6	9.82	57	19	38	1444
82	Tunisia	15.5	10.1	23	20	3	9
75	Sri Lanka	6.3	10.59	13	21	-8	64
80	Syria	18.4	10.81	24	22	2	4
40	Iraq	43.4	11.41	41	23	18	324
71	Slovak Republic	4.8	12.73	11	24	-13	169
9	Belarus	20.3	13.22	26	25	1	1
58	Morocco	44.6	15.31	42	26	16	256
41	Ireland	6.9	18.46	15	27	-12	144
31	Finland	30.4	19.66	33	28	5	25
46	Kazakhstan	267	19.86	58	29	29	841
42	Israel	2.2	21.77	2	30	-28	784
28	Denmark	4.2	21.89	10	31	-21	441
27	Czech Republic	7.7	26.61	16	32	-16	256
63	Peru	128	27.73	53	33	20	400
90	Vietnam	31	27.86	35	34	1	1
67	Romania	23	30.63	27	35	-8	64
85	Ukraine	57.9	30.83	46	36	10	100
36	Greece	12.9	31.47	21	37	-16	256
6	Austria	8.2	33.54	17	38	-21	441
89	Venezuela	88.2	35.27	49	39	10	100

79	Switzerland	4	36.45	8	40	-32	1024
78	Sweden	41	36.61	40	41	-1	1
64	Philippines	29.8	37	32	42	-10	100
10	Belgium	3	40.91	5	43	-38	1444
53	Malaysia	32.9	41.84	36	44	-8	64
62	Pakistan	77.1	46.66	48	45	3	9
29	Egypt	99.5	50.13	50	46	4	16
73	South Africa	121.4	52.84	52	47	5	25
3	Argentina	273.7	64.71	59	48	11	121
60	Netherlands	3.3	70.19	7	49	-42	1764
65	Poland	30.4	75.55	34	50	-16	256
39	Iran	162.9	84.62	55	51	4	16
83	Turkey	76.9	111.46	47	52	-5	25
74	Spain	49.9	147.78	43	53	-10	100
55	Mexico	194.4	164.44	56	54	2	4
43	Italy	29.4	190.86	31	55	-24	576
32	France	54.8	219.41	44	56	-12	144
86	United Kingdom	24.2	223.39	28	57	-29	841
35	Germany	34.9	304.42	38	58	-20	400
38	India	297.3	419.49	60	59	1	1
44	Japan	36.5	430.18	39	60	-21	441

$\Sigma n = 60$

$\Sigma d^2 = 21554$

Using the table above one could draw a scatter diagram for the area hypothesis:



Although there are some clear outliers on the graph, from the line of best fit, there is a clear positive correlation between the area of a country and its GDP. To check that the correlation is positive one can find the gradient of the line of best fit. Using the co-ordinates (300,150) and (170,100) to find the gradient of the line of best fit:

$$\therefore \frac{150-100}{300-170} = \frac{50}{130} = 0.38x$$

$$\frac{y_2 - y_1}{x_2 - x_1}$$

That shows that the line of best fit has a positive gradient and therefore there is a positive correlation between those two variables (area and GDP).

From the scatter diagram and the gradient of the line of best fit, one could see a positive correlation, but still cannot determine the exact strength of correlation.

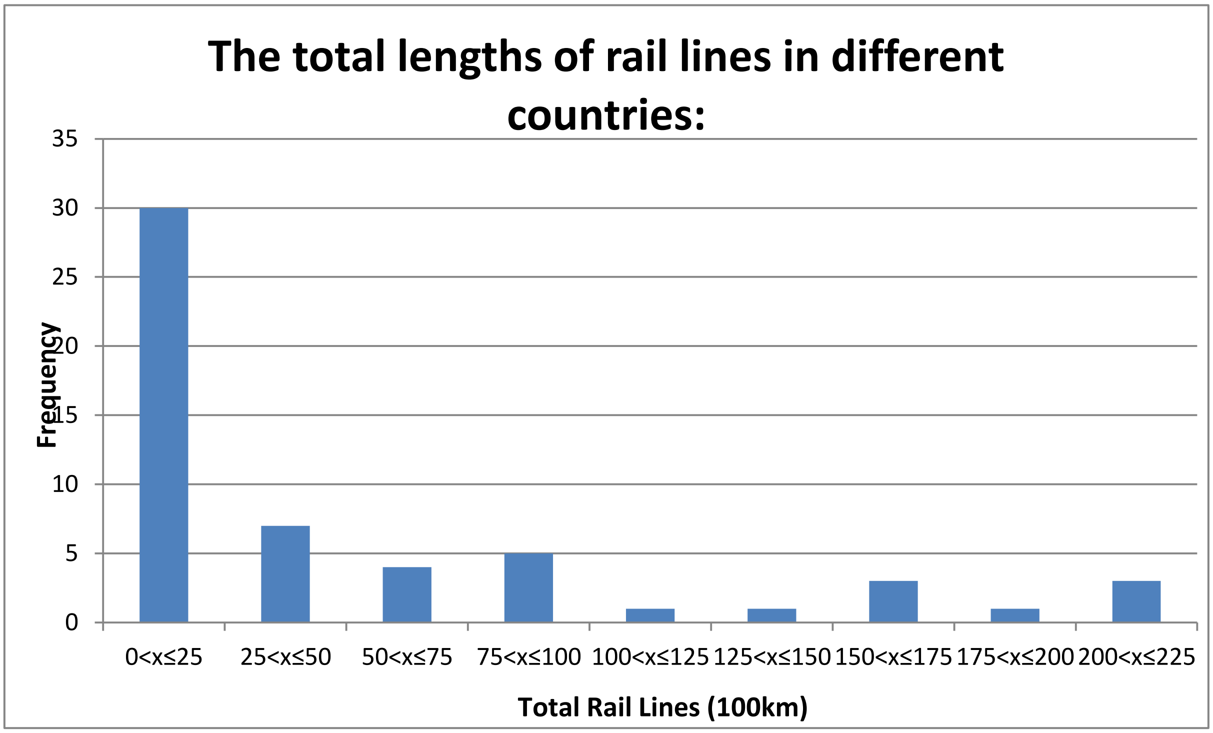
To find to what degree those variables are associated, on can calculate the S.R.C.C:

$$\text{S.R.C.C} = 1 - \left(\frac{6\Sigma d^2}{n(n^2-1)}\right)$$

$$\therefore \text{S.R.C.C} = 1 - \frac{6(21554)}{60(60^2-1)} = 1 - \frac{129324}{60(3599)} = 1 - \left(\frac{129324}{215940}\right)$$

∴ = 1-0.6= **0.4** ∴ **S.R.C.C= 0.4**

That shows that the variables are fairly positively correlated. In comparison to the S.R.C.C obtained from the pilot study on the same variables, the S.R.C.C here is larger (by 0.05). That is because the sample here is larger and therefore the results are more accurate and representative. This supports the second hypothesis because S.R.C.C clearly shows that there is a positive correlation between the area of a country and its GDP.

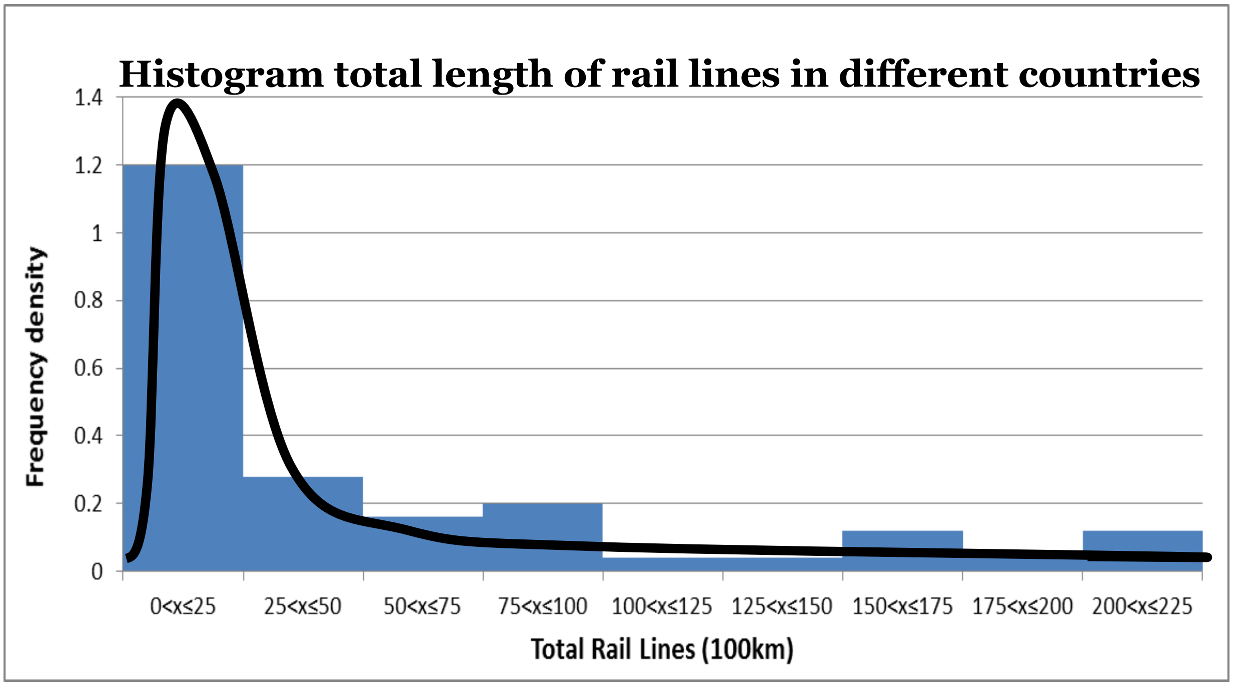


Total Length of Rail Lines

- Finally we will construct a table for the main factor in the third hypothesis which is the total lengths of rail lines in a country. The table will look similar to the ones previously constructed for the first two variables but this time the variable will be the total rail lines:

Total Rail Lines (100km)	Frequency (f)	Class width (cw)	Frequency density (fd)
0<x≤25	30	25	1.2
25<x≤50	7	25	0.28
50<x≤75	4	25	0.16
75<x≤100	5	25	0.2
100<x≤125	1	25	0.04
125<x≤150	1	25	0.04
150<x≤175	3	25	0.12
175<x≤200	1	25	0.04
200<x≤225	3	25	0.12

This table which includes the frequency densities (plotted on the y-axis) will be used to draw a histogram. The whole point of drawing a histogram is to layout the data in a graph which can let the reader visualize the data. The histogram will also show the skewness of the data which will enable the author to view the distribution of the data.



From this histogram above, we can see that there is a positive skew as most of the data is clustered towards the lower end. It shows that most of the countries in the investigation have a small length of rail lines.

Using the histogram above we can identify the probability of a certain interval. To find the probability of an interval you find the area of that interval and divide it by the whole area of all the intervals. After that you multiply the fraction by 100 to get your results as a percentage.

Area of the first interval= 1.2x 2,500= 3,000 km rail lines.

The area of all the intervals= (1.2+0.28+0.16+0.2+0.04+0.04+0.12+0.04+0.12) ×2,500

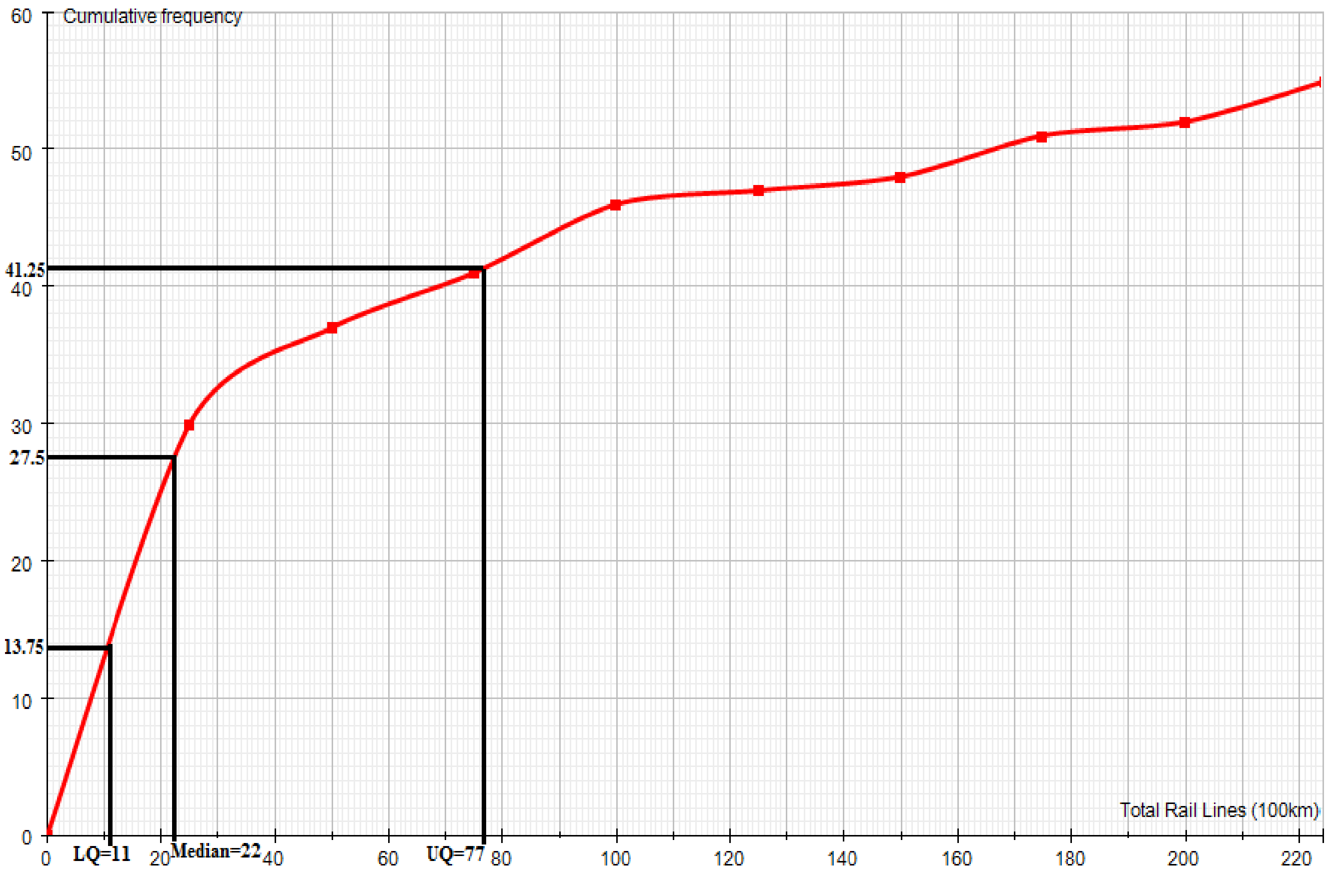
= 2.2 x 2,500= 5,500 km ∴ probability= $\frac{3{,}000}{5{,}500} \times 100 = 55\%$

From finding the probability of the first interval, 55% of the countries in the investigation (excluding outliers) have a total rail line length between 0 and 2,500km. This supports the distribution of the data found from measuring the skewness of the histogram because the histogram is positively skewed which shows that most of the data is clustered at the lower end.

- Constructing a table for the last variable which is the total length of rail lines in a country:

Total Rail Lines (100km)	Frequency(f)	Cumulative frequency (c.f)
0<x≤25	30	30
25<x≤50	7	37
50<x≤75	4	41
75<x≤100	5	46
100<x≤125	1	47
125<x≤150	1	48
150<x≤175	3	51
175<x≤200	1	52
200<x≤225	3	55

This table above will allow the author to plot a cumulative frequency diagram for the third and final variable. Using the upper values of the intervals and the c.f, we can draw a cumulative frequency diagram:



Using this cumulative frequency graph we can calculate many things:

LQ=$\left(\frac{n}{4}\right)^{\text{th}}$ value= $\frac{55}{4}$=13.75th value= 1,100km

Median=$\left(\frac{n}{2}\right)^{\text{th}}$ value= $\frac{58}{2}$=27.5th value= 2,200km

UQ=$\left(\frac{3n}{4}\right)$th value= $\frac{165}{4}$=41.25th value= 7,700km

IQR= UQ-IQ= 7,700-1,100=6,600km

Maximum Value = 22,500km

Minimum Value =280 km

In the beginning, when constructing a frequency distribution table for the total length of rail lines, the median interval was **0<x≤2,500km.** Now from the cumulative frequency diagram, we can see that the exact median is 2,200km, which lies between the median intervals found previously.

- To just make the data more clear, one can summarize the information obtained from the cumulative frequency diagram in a table:

Measures of spread	**Cumulative frequency rail lines (100 km)**
Minimum value	2.8
LQ	11
Median	22
UQ	77
IQR	66
Maximum value	225

Before constructing a box and plot diagram one should first find any outliers so that they can be excluded from the data to make it more representative

To see if there are any outliers:

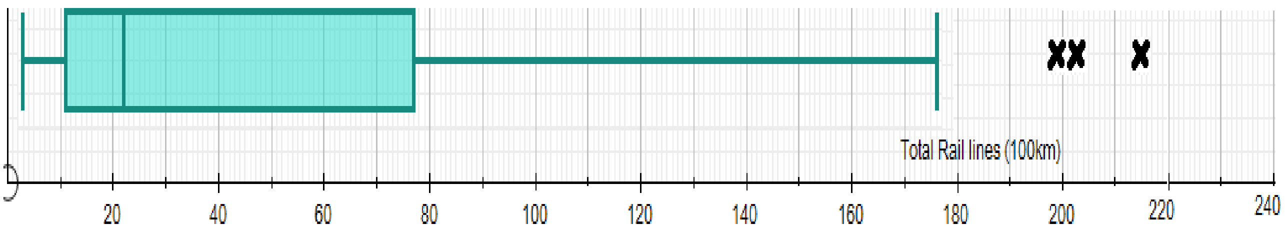
x= IQR × 1.5 = 66×1.5=99

- Highest value not an outlier (H)= upper quartile + x = 77+99=176

Any value greater than 176(>176) is an outlier. Therefore, we will exclude them from the data to make the results more accurate and representative. Listed below are the three outliers that will be excluded:

Country	Total Rail lines (100km)
Poland	197.6
Japan	200.4
Ukraine	216.8

These are the outliers in the data and will be excluded. They will be plotted in the box plot diagram as crosses.



UQ-Media=55 and Median-LQ=11, thus the median is closer to the LQ then to the UQ and that means that the data is positively skewed. This agrees with the skewness found in the histogram for the "rail lines" variable.

- Finally moving to the third and final hypothesis which states, "the higher the GDP of a country, the longer the length of its total rail lines". From the pilot study conducted previously, there was a fairly strong correlation between both variables, so now we will see what the correlation is for the whole sample. Just like the other variables, before drawing a scatter diagram, the author will first construct a table which contains the data which will be used to test the validity of the third hypothesis.

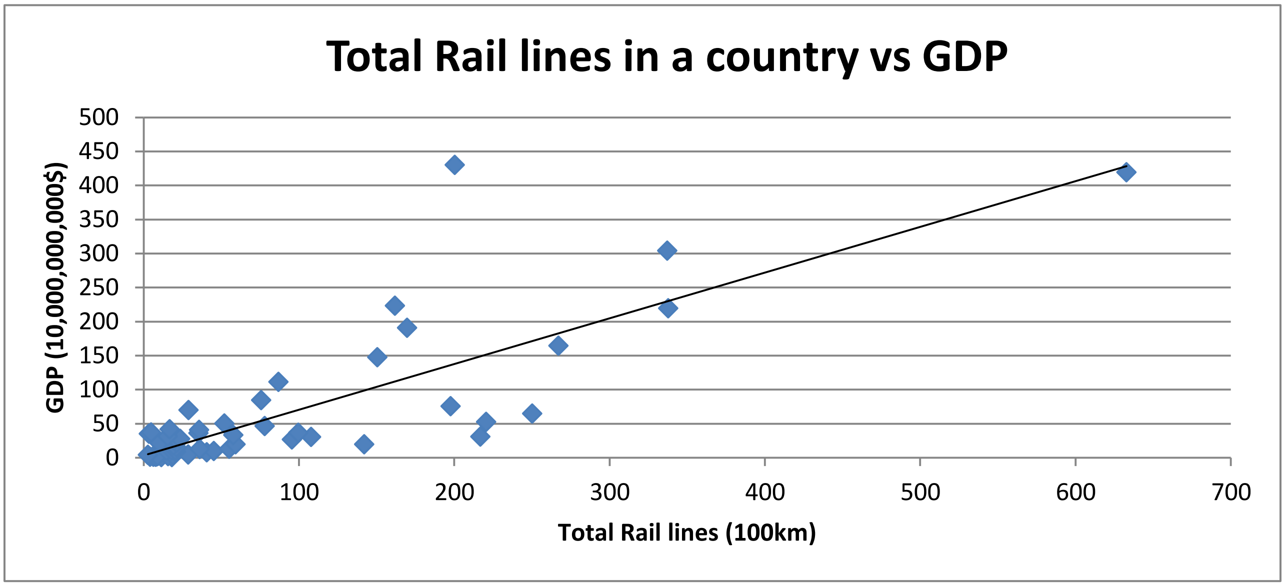
#	Country	Total Rail lines (100km)	GDP (10,000,000,000$)	Rank Rail lines	Rank GDP	d	d²
56	Moldova	11.3	1.11	15	1	14	196
57	Mongolia	18.1	1.12	21	2	19	361
48	Kyrgyz Republic	4.2	1.23	3	3	0	0
11	Benin	7.6	1.4	8	4	4	16
81	Tajikistan	6.2	1.49	5	5	0	0
4	Armenia	8.5	1.69	11	6	5	25
24	Congo, Rep.	8	1.73	9	7	2	4
17	Burkina Faso	6.2	2.18	6	8	-2	4
34	Georgia	15.7	2.26	18	9	9	81
33	Gabon	8.1	2.29	10	10	0	0
52	Macedonia, FYR	7	2.3	7	11	-4	16
30	Estonia	9.3	2.76	13	12	1	1
14	Botswana	8.9	2.79	12	13	-1	1
49	Latvia	18.9	3.66	22	14	8	64
51	Luxembourg	2.8	4.39	1	15	-14	196
12	Bolivia	28.7	4.81	31	16	15	225
70	Serbia	40.6	8.23	36	17	19	361
7	Azerbaijan	20.8	8.99	27	18	9	81
68	Sudan	45.1	9.82	37	19	18	324
82	Tunisia	19.9	10.1	24	20	4	16
75	Sri Lanka	14.6	10.59	16	21	-5	25
80	Syria	18	10.81	20	22	-2	4
40	Iraq	20.3	11.41	26	23	3	9
71	Slovak Republic	36.2	12.73	35	24	11	121
9	Belarus	55.1	13.22	39	25	14	196

58	Morocco	21.1	15.31	28	26	2	4
41	Ireland	19.2	18.46	23	27	-4	16
31	Finland	59.2	19.66	41	28	13	169
46	Kazakhstan	142.1	19.86	48	29	19	361
42	Israel	10.1	21.77	14	30	-16	256
28	Denmark	21.3	21.89	29	31	-2	4
27	Czech Republic	95.4	26.61	45	32	13	169
63	Peru	20.2	27.73	25	33	-8	64
90	Vietnam	23.5	27.86	30	34	-4	16
67	Romania	107.8	30.63	47	35	12	144
85	Ukraine	216.8	30.83	54	36	18	324
36	Greece	15.5	31.47	17	37	-20	400
6	Austria	57.8	33.54	40	38	2	4
89	Venezuela	3.4	35.27	2	39	-37	1369
79	Switzerland	35.4	36.45	33	40	-7	49
78	Sweden	99.5	36.61	46	41	5	25
64	Philippines	4.8	37	4	42	-38	1444
10	Belgium	35.8	40.91	34	43	-9	81
53	Malaysia	16.7	41.84	19	44	-25	625
62	Pakistan	77.9	46.66	43	45	-2	4
29	Egypt	51.9	50.13	38	46	-8	64
73	South Africa	220.5	52.84	55	47	8	64
3	Argentina	250.2	64.71	56	48	8	64
60	Netherlands	28.9	70.19	32	49	-17	289
65	Poland	197.6	75.55	52	50	2	4
39	Iran	75.6	84.62	42	51	-9	81
83	Turkey	86.9	111.46	44	52	-8	64
74	Spain	150.4	147.78	49	53	-4	16
55	Mexico	267	164.44	57	54	3	9
43	Italy	169.6	190.86	51	55	-4	16
32	France	337.8	219.41	59	56	3	9
86	United Kingdom	161.7	223.39	50	57	-7	49
35	Germany	337.1	304.42	58	58	0	0
38	India	632.7	419.49	60	59	1	1
44	Japan	200.4	430.18	53	60	-7	49

Σn= 60

Σd² *=8634*

Now using the table above one will be able to draw a scatter diagram for the hypothesis



which is “the higher the GDP of a country, the longer the length of its total rail lines”:

From the line of best fit we can see that there is a clear positive correlation between the two variables (GDP and length of rail lines), even though there are a few outliers. To mathematically prove that there is a positive gradient, on can do the following:

The two co-ordinates that will be used to find the gradient of the line of best fit are (260,150) and 370,250).

$\frac{y_2 - y_1}{x_2 - x_1}$ $\therefore \frac{250-150}{370-260} = \frac{100}{110} = 0.91x$

That shows that there is a positive correlation between the two variables. To know precisely the correlation between both variables, on can calculate S.R.C.C:

S.R.C.C = $1 - \left(\frac{6\Sigma d^2}{n(n^2-1)}\right)$

$\therefore$ S.R.C.C = $1 - \frac{6(8634)}{60(60^2-1)} = 1 - \frac{51804}{60(3599)} = 1 - \left(\frac{51804}{215940}\right)$

= 1-0.24= **0.76** ∴ **S.R.C.C= 0.76**

S.R.C.C shows that there is a strong correlation between the GDP of a country and the total rail lines in it. The difference between the S.R.C.C obtained for the pilot study for the same variables and this one is 0.16. That is because this is more representative and accurate as the sample size is 4 times bigger than the sample used from the pilot study. The S.R.C.C supports the third hypothesis and shows that there is a clear correlation between the GDP of a country and the total rail lines there.

Conclusion

By carrying out this investigation one was able to prove that the three predictions made initially in were accurate. Through drawing scatter diagrams and finding S.R.C.C one was able to visually and mathematically find out the strength of the correlation between the variables in the hypotheses.

The first hypothesis stated, "the higher the population of a country, the higher its GDP" and to test for feasibility, the author conducted a pilot study on a sample of 15 countries to see if there is a hint of positive correlation or not. By drawing a scatter diagram initially and observing the trend of the data, it looked like the data followed a positive correlation; however to be certain, the S.R.C.C was calculated for the pilot study and it was 0.625 which showed that the 15 samples were fairly positively correlated. That gave the author the initiative to carry out the investigation on a larger sample. After drawing a scatter diagram for the full sample of 60, and although there were a few outliers, the line of best fit on the diagram showed that there was some form of positive correlation. To be more certain, the gradient of the line of best fit was calculated to be certain. The gradient was 4.5x and that made the author assured that there was a positive correlation between those two variables in the first hypothesis. But the gradient did not tell the author exactly how strong the correlation is, so to precisely know the strength of the correlation between the variables, there S.R.C.C was calculated and it turned out to be 0.75 implying that the correlation between the population and the GDP of a country was strongly positively.

The second hypothesis stated, "the bigger the area of a country, the higher its GDP" and by drawing a scatter diagram in the pilot study on a sample of 15 countries, one was able to see that there was a positive correlation. After that, by finding S.R.C.C (0.35) for these 15 data, it was evident there was a weak positive correlation between the variables. This gave the author the incentive to continue carrying out the investigation on the full scale sample which consisted of 60 countries in an attempt to make the results more reliable and representative. A scatter diagram was drawn for the whole sample. Before calculating the S.R.C.C, we found the gradient of the line of best fit just to double check that the graph shows a positive correlation between the variables. The gradient of the line of best fit was 0.38x and that made the author sure that the correlation was positive between the variables. Finally, to measure the strength of the correlation between the variables, the S.R.C.C was calculated and it was 0.4, which was higher than the one obtained from the pilot study (0.35). Although it was carried on a bigger sample (which is more reliable), the S.R.C.C was higher which shows that there is some form of association between both of those variables. This confirmed to the author that there was a moderate positive correlation between the two variables in the second hypothesis.

In the third and final hypothesis, we predicted that "the higher the GDP of a country, the longer the length of its total rail lines". In comparison to the other two hypotheses, the author was not completely sure about this prediction but gave it a try. We used a pilot study which consisted of 15 data to draw a scatter diagram to view the correlation between those two variables. To the astonishment of the author, there seemed to be some sort of correlation between both variables. The S.R.C.C of the data in the pilot study was 0.6 which showed a moderate strong positive correlation. After finding the S.R.C.C, we were very enthusiastic to continue and conduct a full scale investigation to view the results in a more reliable and representative manner. Following the pilot study, we used the whole sample composed of 60 data to draw a scatter diagram to visually view the correlation between the two variables. There seemed to be some sort of positive correlation, but to be certain we found the gradient of the line of best fit which was 0.91x. This showed us that there was a positive correlation between those two variables. The S.R.C.C was 0.76 which signifies a strong linear positive correlation.

By comparing the three hypotheses, there was a positive correlation between all the variables. However, the strength in the correlation varied from one hypothesis to another. Through this investigation, by conducting various statistical techniques, the correlation between the two variables in the third hypothesis were the most positively correlated and that is because their S.R.C.C was 0.76. The second most correlated variables were the ones in the population hypothesis as their S.R.C.C was 0.75 and finally the least correlated variables were the ones in the first hypothesis as their S.R.C.C was 0.4.

Improvements and limitations

In this investigation the author relied on secondary data. An improvement will be to use primary data that is collected by the author, rather than using secondary data collected by others and thus possibly subject to an element of neglect. However, we used a reliable source (World Bank) to try and make the results more representative and unbiased.

The data is limited in that it is restricted to certain countries. An improvement to make the results more reliable and representative will be to increase the sample size and use data that takes into account every single country in this world. Although there were many missing figures for certain countries, an improvement will be to find those figures from other reliable sources and include them in the investigation.

Moreover, a lot of the conclusions formed were based on the S.R.C.C between two variables. One has to note that correlation does not imply causation and as a result, there could have potentially been a third factor which influenced the two variables in question. Thus, it has been a generalization all throughout that the variable in hand is what specifically results in the rise in GDP.

If the author were to repeat this investigation and had more time, a larger sample could have been used along with more sophisticated high demand statistical techniques to reach more precise results.

Overall, this investigation was successful and went very well. However, if it were to be repeated, more data could have been collected from more sources. This would have helped the author reach more representative conclusions. At the moment, the author has evidence that hugely supports the hypotheses, but if a worldwide sample was used (which includes all countries); more varied and representative results would have been generated.

BIBLIOGRAPHY

Abraham, Katherine G. and Christopher Mackie, eds, (2005). Beyond the Market: Designing Nonmarket accounts for the United States, Washington D.C., National AcadethePress.

Blank, Rebecca M. and Mark H. Greenburg (2008). “Improving the Measurement of Poverty, Discussion Paper 2008-17, The Hamilton Project: The Brookings Institution.

Citro, C. and R. Michaels, eds. (1995). Measuring Poverty: A New Approach. Washington D.C.: National Academies Press. Coy, Peter. 2009. “What Good Are Economists Anyway?”

Kuznets, Simon (1941) National Income and Its Composition, 1919-1938. United States Government Printing Office.

Nordhaus, William D., and Edward C. Kokkelenberg, eds. (1999) Natures Numbers: Expanding the National Economic Accounts to Include the Environment. Washington, DC: National AcadethePress.

Nordhaus, William D and James Tobin.(1973) “Is Growth Obsolete.” In The Measurement of Economic and Social Performance, ed. Milton Moss, 509-534. New York: Columbia University Press.

Okubo, Sumiye, Carol A. Robbins, Carol E. Moylan, Brian K. Sliker, Laura I. Shultz and Lisa S. Mataloni. (2006). “BEA’s 2006 Research and Development Satellite Account,” Survey of Current Business 86: 14-27.

Stiglitz, Joseph E. (2009). “GDP Fetishism,” The Economist Voice. Retrieved from: www.bepress.com/cgi/viewcontent.cgi?article=1651&context=ev.

Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. (2009) Report by the Commission on the Measurement of Economic Performance and Social Progress.

Picketty, Thomas and Emmanuel Saez. (2009) “Incothe author Inequality in the United States, 1998-2002”