Sheffield Hallam University | College of Business, Technology and Engineering

MSc Dissertation Report

"TACKLING THE URBANISATION AND POPULATION CRISIS IN PAKISTAN USING BIG DATA ANALYTICS"

A dissertation submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Master of Science in Big Data Analytics

Student Name	Tallal Ahmed Bhatti
Student ID	31042155
Supervisor	Dani Papamaximou
Date of Submission	15th September 2022

This dissertation does NOT contain confidential material and thus can be made available to staff and students via the library.

Acknowledgements:

Firstly, I would like to thank Allah (S. W. T) for being a source of guidance and blessing me with the strength to complete the research as part of my academic programme.

A special thanks to all the teachers at Sheffield Hallam University who have been forever helpful which has made the journey easier. I am forever grateful to my supervisor, Dani Papamaximou, for guiding me through the entire process and making me work harder than ever. All your guidance and support has helped me a lot and it was great being your student and grasp great amount of knowledge from you.

To my lovely parents who made my dream of pursuing a postgraduate degree as an International Student come true. For always believing in my strengths and abilities for which I am forever indebted to them. My siblings, Warda, Bilal, Farwa, Aadil, Zara, Ali, and Tayyab for motivating me whenever things went difficult.

Last but not the least, to the friends I made in this new country and the ones back home. Thank you for sharing new ideas with me and forever supporting me.

To all the above-mentioned people, this research was a great journey because of you which helped me learn a lot.

Abstract:

The following research aims to investigate the population and urbanisation crisis in Pakistan. The research was conducted to investigate the factors that influence in the urbanisation and population increase within a few districts in Pakistan. The past research papers mostly focus on the issues caused by overpopulation and there is little to no research as to why there are only a handful of districts that constitute a large portion of population. The research aims to explore the key variables driving the rapid increase in population and urbanisation in Pakistan by discussing unfair distribution of population and urban areas across the country. Increase in Urbanisation and population are not harmful if they are fairly distributed across the country which is not the case in current research topic. Data collected from different sources was joined together to perform analysis and create a machine learning model which helps identifying variables that are most influential towards the research problems.

Contents

1.Introduction

How can development of less-developed districts help towards tackling increase of Urbanisation and Population crisis in Pakistan by determining key factors constituting towards it?

1.1 Project Rationale

One of the major problems faced by Pakistan since its birth has been the lack of fair distribution of urban cities and overpopulation in metropolitan cities which has led to impacts that can be dreadful for the country in the coming future if not dealt with accordingly. With the ever-increasing population in Pakistan, the few metropolitan cities such as Lahore, Islamabad and Karachi seem to have increased in size over time to accommodate its inhabitants. The overcrowding in such metropolitan cities has led to increased unemployment, pollution in the cities, traffic, and most importantly housing crisis. Considering recent years, the district of Lahore has grown from 263.51 square kilometers in 2000 to 426.8 square kilometers in 2015 almost doubling the previous figure (GHS, 2018). The population in Lahore has also increased from 8.3 million in 2000 to a massive 11.1 million in 2015 (GHS, 2018).

Rapid urbanisation in Pakistan has become more of a problem than a solution to its problems. Currently, one-third of Pakistani population lives in Urban areas and will increase to half of the country's population in 2030 (Kugelman, 2013). The massive void which is created between these developed metro-cities and the rest of the cities can be addressed by planning new urban areas using key factors that influence the rapid increase of urbanization in Pakistan. This paper aims to tackle the Urbanisation and population crisis in Pakistan by identifying key components and their effects using Big Data Analytics.

1.2 Project Scope

The scope of the project focuses on developing a regression model and a report which will determine the most influential factors that contribute the most towards rapid urbanization in Pakistan. Most of the past research regarding the population of Pakistan revolve around predicting population and other variables whereas a little too few only discuss or investigate the factors associated with it (Kugelman, 2013). The dataset collected for the following research also has strong multicollinearity amongst its variables which makes it difficult to make predictive models. The variables have strong multicollinearity because most of them are driven from each other. The key variables and their relationship with one another are investigated from the initial stages of the research till the final stages of the research.

1.3 Project Aims

The following research aims are crafted based on the initial research questions.

- Identify the districts that contribute the most towards population and urbanization
- Identify all the initial variables that contribute towards ranking the development of districts.
- Identify the relationship between variables and the strength of it
- Develop a Machine Learning Model that identifies the most important variables contributing towards the crisis
- Develop a report using the model that answers the business questions of the following research

The aim of the research is to explore the factors associated with population and urbanization by identifying relationships between them and how strong these relationships are to one another. The goal is to develop a statistical model, using Machine Learning algorithms, which will help us identify the most effective variables that are causing rapid urbanization in Pakistan. The other aims of the research are to visualize the outputs determined by the total project. The final report will include solutions to the problem and how it allows future research in the given topic.

1.4 Project Objectives

The objectives for the following research are made to ensure that the project aims are realized:

- Complete a literature review using past research papers, articles, etc. to understand current research
- Gather data from concerned organisations regarding the given research
- Identify key variables and factors in the dataset to be used in further research
- Create statistical models using Machine Learning
- Measure accuracy of the model to get valid results
- Test the system for anomalies
- Draw visualizations to answer relevant questions for the research
- Evaluate results and draw future scope for the research

The initial objective of the research is to develop an understanding of the topic and the issues addressed by conducting literature review of the given topic. The literature review would include references from past research work, journal publications, etc. which will help in identifying key issues and variables associated with them. After completing the literature review, the relevant data for the research is collected from different sources such as government organisations and NGOs. The data is cleaned and made valid to identify key variables and their relation to one another. Once the relationships between variables and their strength is investigated, a model will be developed as per the need of the research which will explain the

variables in more detail (Rong & Bao-wen, 2018). After selecting a suitable model for the research, the results discovered are again trained and tested to avoid anomalies and make the model more accurate (Ben Braiek & Khomh, 2020). Furthermore, these results are visualised on a broader scale to analyse the research in a different perspective and answer relevant questions of research (Kirk, 2016). Finally, a conclusion will be drawn with respect to the research conducted and future scope of the research will be defined.

1.5 Project Benefits

The following research is beneficial in many terms as it can help identify the patterns in which urbanization is rapidly increasing in Pakistan. It also discovers key factors that play the most important role in urbanization and overpopulation. The following factors can be used to determine solutions for the issues at hand. Urbanisation is a great form of growth but uneven urbanization and that too at a rapid speed can be devastating. Such is the case with Pakistan as the urban cities are growing at an alarming rate and citizens must go through difficulties such as Unemployment, Health Issue, Education, etc. (Shaikh & Nabi, 2017). The following research will help in eliminating such issues with the use of big data analytics.

1.6 Achieving the Research Objectives

The initial proposal of the research can be found in the appendix section (Appendix A) where literature review was conducted along with research design and methodology. The ethics and usage of concerned data is also discussed in the attached proposal. The approval to use this data was acquired before the research was conducted on it. The literature review in the proposal discusses the reasons to choose certain areas/variables for our analysis. The research initiates by completing a literature review on the given topic and selecting appropriate variables for our research. Research methodology follows the next step where the data is classified as primary or secondary data. In the following research, the data used is a secondary data as it was collected by some other organisations that have given the permission to use for further research and analysis which can be found in Appendix Section (Appendix B). Then comes the model development stage of the research where statistical models are created by using machine learning techniques. The model is trained and tested to give final variables that would suit the research questions the most (Ben Braiek & Khomh, 2020). The efficiency of the model is calculated and improved until a decent value is achieved and the model can deem as successful. The insights and variables provided by model are visualised to deliver better insights and information which would further help in the future for research purposes. The research concludes with a conclusion summary to help eliminate the issues discussed in the research.

2. Literature Review

For the following research being conducted, the initial key literature was conducted to discover key factors that play the most important role in urbanization and population of a country. After the initial variables for the research are determined, the data is collected from different sources. In current research, the data set mainly comes from the population census conducted by the Pakistan Bureau of Statistics in 2017 and 1998. Furthermore, medical facilities dataset is also used to support the census data and is provided by Al-Hasan Systems through its open data pioneering initiative. The datasets collected from the provided sources are free to use for research purposes and have been ethically approved before the research was conducted. Using these initial variables, the data is cleaned to give one clear output dataframe containing all necessary variables. The next stage is selected of correct algorithms to develop a model which answers the research question (Nogueira, Brown, & Sechidis, 2018). The selection of algorithms, techniques and models should be carefully completed to avoid discrepancies in the future of the research. Table 2.1 explains how the literature review is divided into further sections which explain the entire literature process of the research.

Table 2.1 Division and explanation of Literature Review

Literature Area	**Explanations**
Variable Selection	• Discover variables that constitute towards the increasing urbanization and populations • Identify the need to choose these variables and how they are beneficial to be included in the model
Statistical Technique Selection	• Explain the selection of the statistical techniques used to conduct the research and generate the model • Validate choice of choosing the technique and related algorithms • Identify what can be achieved using the following technique
Model Implementation	• Discover the best approach towards creating a stable and accurate model • Identify the insights and variables provided by the model • Investigate multicollinearity in variables and how to tackle it • Summarise the model description in analytical terms
Test	• Identify if the system model is accurate and what steps can increase the accuracy • Identify how the final model answers the research questions

2.1 Variable Selection

In every research it is vital that the key variables are discovered before the algorithms are implemented to develop a statistical model (Kuo & Mallick, 1998). The focus is to collect data for variables and then explain what the intentions are to use it for.

2.1.1 Data Collection

To create the dataset, the variables needed to be selected after carefully completing the literature review. The dataset is collected from Pakistan Bureau of Statistics where the data for population and its associated variables is available for research and academic purposes. According to United Nation's Development Program, Pakistan has the highest urbanization rate in South Asia and estimates that half of the country's population will be living in urban cities by 2030 (Bari, 2020). The housing situation in urban cities of Pakistan is a gigantic issue that needs to be addressed in the following research. It is predicted that by 2030, Pakistan's five largest cities will account for 78% of shortage in housing units. Furthermore, the lack of quality education and health facilities help in increasing the urbanization rate in Pakistan (Shaikh & Nabi, 2017). Hence, the data regarding urbanization and population was collected from the Pakistan Bureau of Statistics which included data regarding housing, employment, education, and the data for health facilities was acquired from Al-Hasan Systems. The data provided by concerned organisations is mostly at district level hence the following research is also conducted on district level.

2.1.2 Intended use of variables

The goal of the research is to study the pattern of urbanization and population in Pakistan so that the variables constituting towards it can be identified and used in such a way that districts with lesser development can be developed. These variables will be further used in the research to identify the type of relationship they have to one another and how one variable effect the other (Gogtay & Thatte, 2017). Since the research intends to address the population and urbanization issues in Pakistan, it would be beneficial to use total population as a target variable around which the whole research model will be generated. Housing Units in urban and rural areas, division of students, employed and unemployed personnel and medical facilities were considered for our concerned research. The provided dataset had a very complex structure and needed to be restructured to get the desired variables in an appropriate format. The model generated will help decide which variables have the highest coefficient in an accurate model which would further help in deciding how to plan development in districts and what are the key variables adding to it (Poole & O'Farrell, 1971).

Taking all this into account, the final dataset would have all the necessary variables required to conduct the following research. Some variables and tables were not considered when collecting and finalizing data because they are not relevant enough to the research. For example, number

of rooms in a housing unit, fuel type in housing units is irrelevant because the research would mainly focus on population and urbanization which means a sole focus towards urban areas. Since urban areas do not really have an access of fuel or water problem, such variables can be considered negligible to our research.

2.2 Algorithm Selection

Selecting the right algorithm is one of the most important tasks while conducting the research. Selecting a wrong algorithm would lead to generating a false and inaccurate model which would result in invalid research (Lee & Jae Shin, 2020). Before selecting the right algorithm for the research, it is important to understand the data collected and how it is related to each other. Hence, after analysing the situation in the following research it is beneficial to use Linear Regression model as the pair plot drawn after final data frame is achieve shows linear relationship amongst the variables. The Ordinary Least Squares Regression method is used as the research was supposed to consider unknown parameters as well. OLS is used in the following research because OLS estimators have the least variance among all linear and unbiased estimators (T. Pohlmann & W. Leitner, 2003). Principal Component Analysis is another technique used to cluster variables with high multicollinearity into respective principal components to increase the efficiency of the generated model (Abdi & J. Williams, 2010).

2.2.1 Correlation and Regression

Having achieved the final fact table containing all necessary variables for our research, the next part of the research analysis begins with the exploratory data analysis. The initial stage of exploratory data analysis in the following research is determining the correlation between the variables to select the right statistical technique for the following analysis. Correlation function is used on the final data set to calculate the correlation values between the variables which define whether the variables are correlated to each other or not (Gogtay & Thatte, 2017). A correlation matrix is generated which is further investigated to select appropriate statistical techniques. The correlation matrix is further visualised in heatmap using different libraries such as seaborn and matplotlib to understand the relations between variables in a broader perspective. Once the correlation is achieved, the exploratory analysis continues by plotting pair-plots between the variables to check and confirm which type of statistical method is required in the following research. As per the results of the pair-plot graphs, a regression technique will be used on the data set. Choosing and applying the correct method of regression on the final dataset is the most important part for the research to be considered accurate and valid (Raheem, Udoh, & Gbolahan, 2019).

2.2.2 Ordinary Least Squares Regression

Ordinary least square regression is considered one of the most successful regression techniques

which allows to minimize the prediction error between the real and predicted values (Hutcheson & Moutinho, 2011). It uses the total sum of squared values rather than just using values which results in the output not being equal to zero. Using the following technique, the research can continue successfully by minimizing the prediction error and keeping the model accurate and valid.

2.2.3 Principal Component Analysis

One of the more effective techniques in Data Analysis includes that of Principal Component Analysis. Principal Component Analysis is a technique used to interpret the information from large amounts of data sets by reducing the dimensionality of the following datasets (Joliffe T & Cadima, 2016). It creates one or more resultant components which minimize the information loss, reduces eigenvalue concerns and these components successfully maximise variance. Putting it in an easier way, Principal Component Analysis will group variables with high multicollinearity into resultant components which would contain necessary information from all these variables without losing any data and successfully maximise variance. Using Principal Component Analysis in the research model would increase the accuracy of the model and help identify key variables in a broader context (Abdi & J. Williams, 2010).

2.2.4 Algorithm Summary

The flow of work in exploratory data analysis starts with creation of correlation matrix to analyse the variables and their relationship with one another on a broader scale. The strength of correlation will further aid in plotting pair-plots for every variable and their relationship with other variables. Once the pair-plot is achieved, the correct technique is to be applied on our data set. Linear Regression is one technique to identify linear relationship between two variables and OLS is a regression method which minimizes the prediction error keeping the model accurate. Furthermore, PCA would be conducted as per requirements to improve the accuracy and efficiency of the final model (Joliffe T & Cadima, 2016). Finally, the model will be trained and tested before the final model is published to explain the analysis of the research.

2.3 Implementation of Model

The model generated by applying appropriate algorithms to our final dataset. The data is split into train set and test set so that the performance of the machine learning can be measured to predict variables/values. The train dataset is used to fit the machine learning model whereas the test dataset is used to evaluate the fit of the machine learning model (Darlington & Hayes, 2017). Mind that the data used to evaluate the performance of the machine learning model is the new data not the data used in the model training. One of the conditions of using a train-test split is that the data should be large enough to be divided into train and test data. Small datasets cannot be used to because there will not be enough data left in training dataset to test effective mapping of inputs to outputs. The data is divided into train and test data using random selection

to ensure fair and valid representation of original dataset.

2.4 Literature Review Summary:

The literature review for the following research is completed by collecting data and selecting appropriate variables after applying correlation matrix. Before the correlation matrix is obtained, it is necessary to use literature review and human intuition to select initial variables that suit the following research topic. Once the correlation matrix is obtained, pair-plotting of variables is to be conducted to interpret the relationship between the variables. Moving on, an appropriate statistical technique is applied to these variables to develop a machine learning model for the following research. Variables with high multicollinearity are joined together using Principal Component Analysis which will increase the accuracy and validity of the model (Mansfield & Helms, 1982). Lastly, the model is generated by splitting the data into train data and test data using train-test split technique. The model will further provide information on key variables that answer the research questions.

3. Research Methodology

The research was designed to achieve a final report which discusses the key variables that influence the rapid urbanization and population in Pakistan. The research starts with literature review and data collection. After collecting the data, all the necessary steps to clean and validate the data take place. Once the data is clean and ready to use, it is merged to form one final data frame containing all suitable values for the conducted research. The literature review helps in eliminating invalid and unnecessary variables to be used in the research. The data obtained from the data sources is quantitative and can be used for the research. The cleaning process included restructuring and dealing with missing values to ensure fair data throughout the data frame. Further, the influential variables are determined from the dataset using machine learning models (Desboulets, 2018). Obtain and select the most accurate model for the research by applying different analytical techniques (Kuo & Mallick, 1998). The obtained variables from the model will be visualised in different perspectives to answer the research questions.

3.1 Research Ethics

Following the guidelines set for the research, the ethics were practiced for the following research. All the data was collected from official channels responsible for census and all the health data is collected from a private organization (Al-Hasan Systems) which provides open-source data for research and academic purposes. Since all the data collected from these sources come from open-source data, it was easy to use these datasets in the research without any need to acquire authorization to use (Morrow, Boddy, & Lamb, 2014).

3.2 Secondary Research Data

The following research only contains secondary research data as all the data used in the research is collected by some other organisations. The collected data is used to craft data frames corresponding to different variables of the research selected initially through the literature review and intuition (Wazir & Goujon, 2019). The data set obtained from the Pakistan Bureau of Statistics contains the employment, population, education, and housing data set on district level but had to be restructured due to complex nature of the datasets. It was mandatory to analyse the data on a mutual granularity level present amongst all data frames so that joining of data sets is smooth and has valid entries. Every data frame generated for each factor is analysed specifically and separately where irrelevant variables in terms of research area are dropped. The data is cleaned for the final datasets generated for these factors and merged to form one final table containing all variables with district names. Some variables were calculated using formulas to calculate density per area.

3.3 Selected Dataset

The selected data set is obtained from the census report conducted by PBS and Health Facilities data provided by Al-Hasan systems. 22 columns and 117 rows were obtained in the final fact table as the data only contained districts of the country. The federal agency regions and federally administered regions are removed from the final data set because the federally administered regions were given the autonomous powers in respective provinces after the census was conducted (Hashim, 2018). Naming convention needed to be dealt with carefully as the districts are named in Urdu but when written in English can have different spellings which can later cause issues in merging data frames together. The selected dataset also had spatial files or shape files (.shp extension files) to visualize the data on the country's map using BI tools such as Tableau or Pandas library such as geopandas. In the following research, the naming is corrected again in the final table to make sure that it can be joined with data in spatial files.

3.4 System Testing

The variables used in the following research had already been established via secondary research, the literature chose the algorithms as the critical area for evaluating and confirming the prediction accuracy. To test the algorithms, accuracy values were obtained using Scikit-Learn (VanderPlas, 2016). The system is tested to evaluate the accuracy of the model generated. Different factors will help compute the accuracy of model such as Skewness, R-Squared Value, Covariant Coefficient, etc. The accuracy and precision of models will be initially determined using all variables and then different techniques are to be applied to achieve a decent accuracy score for our model. Principal Component Analysis and Ordinary Least Square Regression help in identifying the variables and determining the accuracy of the model (Darlington & Hayes, 2017).

4. Model Development

4.1 Pre-Processing Data

To make sure the data is valid and fair without any missing values, all the data is cleaned accordingly, and the missing values are dealt according to their needs (Famili, Shen, Weber, & Simoudis, 1997). After conducting the literature review, was decided that the research would be conducted using Pandas in Python with use of libraries such as SciKit-learn for machine learning algorithms and seaborn and matplotlib for visualization of data in terms of graphs and tables. The data cleaning process also includes the correction of naming convention in the datasets. This would ensure smooth analysis without any bias in output or models. The data frames are generated in Pandas by reading Excel files and converting them to data frames. Each factor has its own clean data set with variables accumulating towards that factor. For example, the employment data contains the data for unemployed and employed population in a district whereas housing would include housing units in urban areas and rural areas. The data frames constructed are then merged on Districts so that one final table can be achieved which holds all the values required to conduct valid research. The data regarding medical facilities comes from a different source with a different naming convention. A few federal areas and sub districts have also been added as districts in the data set which can cause confusion around our research. To eliminate this, the naming convention had to be restructured and the data frame should be merged using left join with the datasets provided by Pakistan Bureau of Statistics as it is an official government organization with official records and statistics. The purpose of this is to make sure we get data for all the rows in the datasets provided by the Pakistan Bureau of Statistics.

The questions surrounding the research are constructed towards reaching the research goals. These questions are crafted to answer the research questions in terms of all the areas discussed in literature review acting as our variables. The data is queried for high numbers because the research revolves around the idea of tackling the population and urbanisation crisis which means that the research focus should be around high numbers. However, in certain cases the research would require querying data for lowest numbers as well because it would need to understand the differences in two extremes of our dataset (Adeel & Khan, 2017). Another reason to explore higher numbers is the low urbanisation in the country. Since Pakistan is a country with a high rural population, it is fair to avoid using low numbers for our analysis (Alvi, 2018).

4.2 Dataset training and testing

Python is the used for the following research because of its great capability in data analytics with use of libraries such as Numpy, Pandas and SciKit-Learn (Stancin & Jovic, 2019). When it

comes to training and testing the model for our research, the scikit-learn library is imported, and its functions are used to generate an accurate and valid model. All the statistical techniques such as regression, pair plot, Principal Component analysis, exploration of skewness of dataset, etc. are carried out using the scikit-learn library and the functions associated with it.
The variables are identified using the model generated. Initially, there is a large multicollinearity between variables which needs to be eliminated. Furthermore, there can be some outliers in the conducted research and would need to be addressed according to the situation. One way of tackling outliers is by replacing them with median value so that our research remains fair and unbiased. It is important to note that the research would be invalid without dealing with outliers (Stevens, 1984). Hence, it is an essential part of dataset testing and training to deal with the outliers. After dealing with outliers, the target variable is selected (Population for the current research since the research revolves around population). The model is generated initially to get an idea of how accurate it is with respect to the target variable. Values with lowest ViF values shall be selected (less than 10 ViF value in the following case) (Kim, Aloe, & Becker, 2017). The rest of the variables with high ViF values will be clustered into principal components using the principal component analysis. This would provide with a more accurate and suitable R-Squared value for our research model. Following is the flow of how the dataset training and testing will take place:
INSERT THE FLOW DIAGRAM OF TRAINING AND TESTING DATA HERE

5. Final Analysis and Interpretation

The analysis and interpretation of the following research is conducted once all the data in the final tables is cleaned and pre-processed. Before moving on to analysis, it is mandatory to perform data validity checks to make sure all data is valid and there are no missing values.

5.1 Creation and Explanation of Dataframes

Since all the data (except health data) comes from the same data source, it was achievable to create dataframes corresponding to each factor of our research which includes Education, Employment, Housing, and Population dataset (Chen, 2017). The health dataframe was also created the same way as the rest with the only difference being the source of data and naming convention which had to be corrected to merge data successfully. These data frames were then exported as .csv (comma separated values) file using pandas .to_csv function. Creating csv files from these frames helps in preserving their structure and gives a copy for the research without any possibility of losing data. The csv files created after cleaning and pre-processing and valid to use and only contain information regarding their area of concern. For instance, health csv file would have health data only and vice versa.

```
import pandas as pd
import numpy as np
import seaborn as sb
import matplotlib.pyplot as mp
import scipy
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt

df_population = pd.read_csv("PopulationFinal")

df_employment = pd.read_csv("EmploymentFinal")

df_education = pd.read_csv("EducationClean1")

df_housing = pd.read_csv("Housing")


df_health = pd.read_csv("HealthFinal")
```

Fig 5.1 – Generating Dataframes from the csv files created after cleaning data

Observing fig 5.1, all necessary libraries such as numpy, pandas, seaborn, maplotlib, and scipy are imported due to the scope of the research. Data frames are created by reading csv files using the Pandas library for further analysis (Chen, 2017).

In the figure attached below (Fig 5.2), the population dataset is being used in which the district column has naming issues. It is important to remove the ‘DISTRICT’ from the name of districts in District column to make sure that the data can later merge smoothly with health data without the use of complex fuzzy words match function. Furthermore, this is the main dataframe that will be used and all other dataframes will be merged on it as population dataframe includes all the necessary districts and their values in the population dataframe. A prime example of removing the ‘DISTRICT’ from the names in district column would result ‘BANNU DISTRICT’ to become ‘BANNU’. The same process is carried out with all the other districts as well. The population dataframe has 10 columns with their data types listed in the figure below.

```
df_population.dtypes
df_population['DISTRICT'] =df_population['DISTRICT'].str.replace(' DISTRICT','')

df_population
```

```
DISTRICT                                object
AREA (SQ. KM.)                           int64
TOTAL POPULATION                         int64
POPULATION DENSITY PER SQ. KM.         float64
AVERAGE HOUSEHOLD SIZE                 float64
POPULATION 1998                          int64
1998-2017 AVERAGE ANNUAL GROWTH RATE   float64
URBAN POPULATION                         int64
RURAL POPULATION                         int64
dtype: object
```

	DISTRICT	AREA (SQ. KM.)	TOTAL POPULATION	POPULATION DENSITY PER SQ. KM.	AVERAGE HOUSEHOLD SIZE	POPULATION 1998	1998-2017 AVERAGE ANNUAL GROWTH RATE	URBAN POPULATION	RURAL POPULATION
0	BANNU	1227	1167071	951.1600	9.34	675667	2.91	49948	1117123
1	LAKKI MARWAT	3164	875744	276.7800	8.71	490025	3.10	89252	786492
2	DERA ISMAIL KHAN	7326	1625088	221.8200	7.96	852995	3.44	360218	1264870
3	KOHISTAN	7492	784711	104.7400	7.84	472570	2.70	0	784711
4	TORGHAR	454	171349	377.4200	6.49	174682	-0.10	0	171349
5	MANSEHRA	4125	1555742	377.1500	6.51	978157	2.47	144898	1410844

Fig 5.2 - Final dataframe of population dataset with datatypes

Loading the data from the csv file and creating dataframes is the latter part of the research analysis as the csv files were made using the clean dataframes made in the initial processes. After the new dataframes are created, the next objective is to merge them all together so that they can be analysed altogether. In the following research, mentioned in previous chapters, the data will be merged on districts since it is the variable which is constant in every dataset collected for the investigation. Moreover, the research aims to tackle the urbanization and population crisis in Pakistan which means that Population would be the target variable and all the dataframes shall be joined on it. This condition results in applying left join on population dataframe so that none of population data is lost.

```
df_test1 = df_population.merge(df_employment,how='left',on=['DISTRICT'])
df_test1
```

	DISTRICT	AREA (SQ. KM.)	TOTAL POPULATION	POPULATION DENSITY PER SQ. KM.	AVERAGE HOUSEHOLD SIZE	POPULATION 1998	1998-2017 AVERAGE ANNUAL GROWTH RATE	URBAN POPULATION	RURAL POPULATION	WORKED (INCLUDED UN PAID FAMILY WORKER)	SEEKING WORK	STUDENT
0	BANNU	1227	1167071	951.16	9.34	675667	2.91	49948	1117123	153252	30973	163130
1	LAKKI MARWAT	3164	875744	276.78	8.71	490025	3.10	89252	786492	122182	21728	116462
2	DERA ISMAIL KHAN	7326	1625088	221.82	7.96	852995	3.44	360218	1264870	314317	25058	211672
3	KOHISTAN	7492	784711	104.74	7.84	472570	2.70	0	784711	132638	55409	28420
4	TORGHAR	454	171349	377.42	6.49	174682	-0.10	0	171349	25347	4160	15932
...	...	...	...	...	...	...	...	...	...	...	...	...
113	LORALAI	8018	397423	49.57	7.00	250147	2.46	64891	332532	54039	16852	39284
114	MUSAKHEL	5728	167243	29.20	6.78	134056	1.17	14135	153108	19935	7964	10977
115	SHERANI	4310	152952	35.49	7.30	81684	3.35	0	152952	9477	6439	10776
116	ZHOB	15987	310354	19.41	6.63	193458	2.51	46164	264190	33745	14366	28005
117	ISLAMABAD	906	2003368	2211.22	5.86	805235	4.90	1009003	994365	486288	29663	408592

Fig 5.3 - Merged data of population and employment

The following figure (fig 5.3) shows the merged data between Population and Employment data. The merged dataframe is named as df_test1 which has 118 rows and 12 columns. There are some values having zero values in columns such as urban or rural populations because some districts have zero urban/rural areas as per data collected in the census.

Moving on, the next dataframe to be joined with this resultant dataframe (*df_test1*) is the housing dataframe (*df_housing*). Once again, a left join is performed to join the dataframes and explore their datatypes which shows the final columns in the resultant dataframe (Chen, 2017). Figure 5.4 shows the column names and datatypes of resultant dataframe (*df_test2*).

```
DISTRICT                                          object
AREA (SQ. KM.)                                     int64
TOTAL POPULATION                                   int64
POPULATION DENSITY PER SQ. KM.                   float64
AVERAGE HOUSEHOLD SIZE                           float64
POPULATION 1998                                    int64
1998-2017 AVERAGE ANNUAL GROWTH RATE             float64
URBAN POPULATION                                   int64
RURAL POPULATION                                   int64
WORKED (INCLUDED UN PAID FAMILY WORKER)            int64
SEEKING WORK                                       int64
STUDENT                                            int64
OVERALL HOUSING UNITS                              int64
URBAN HOUSING UNITS                                int64
RURAL HOUSING UNITS                                int64
dtype: object
```

Fig 5.4 – data types for merged data after housing is joined (df_test2)

Similarly, same steps will be followed to join the rest of the dataframes. Education is the next dataframe joined to df_test2 and the resultant dataframe (df_test3) is created which now constitutes of education, housing, employment, and population data. Fig 5.5 shows the dataframe created after joining education data. This dataframe is later renamed as df_final to avoid naming issues.

```
df_test3 = df_test2.merge(df_education,how='left',on='DISTRICT')
```

```
df_test3
```

DISTRICT	AREA (SQ. KM.)	TOTAL POPULATION	POPULATION DENSITY PER SQ. KM.	AVERAGE HOUSEHOLD SIZE	POPULATION 1998	1998-2017 AVERAGE ANNUAL GROWTH RATE	URBAN POPULATION	RURAL POPULATION	WORKED (INCLUDED UN PAID FAMILY WORKER)	...	STUDENT	OVERALL HOUSING UNITS	H
BANNU	1227	1167071	951.16	9.34	675667	2.91	49948	1117123	153252	...	163130	118539	
LAKKI MARWAT	3164	875744	276.78	8.71	490025	3.10	89252	786492	122182	...	116462	97527	
DERA ISMAIL KHAN	7326	1625088	221.82	7.96	852995	3.44	360218	1264870	314317	...	211672	197903	
KOHISTAN	7492	784711	104.74	7.84	472570	2.70	0	784711	132638	...	28420	99413	
TORGHAR	454	171349	377.42	6.49	174682	-0.10	0	171349	25347	...	15932	26281	
...	...	...	...	...	...	...	...	...	...	...	...	...	
LORALAI	8018	397423	49.57	7.00	250147	2.46	64891	332532	54039	...	39284	54026	
USAKHEL	5728	167243	29.20	6.78	134056	1.17	14135	153108	19935	...	10977	24325	
SHERANI	4310	152952	35.49	7.30	81684	3.35	0	152952	9477	...	10776	20307	
ZHOB	15987	310354	19.41	6.63	193458	2.51	46164	264190	33745	...	28005	44839	
LAMABAD	906	2003368	2211.22	5.86	805235	4.90	1009003	994365	486288	...	408592	332145	

Fig 5.5 – Education, Employment, Housing, and Population dataframe (df_test3)

The last part to obtain the final fact table containing all the dataframes joined together will be completed after the health data is joined with df_test3. The final dataframe is named as df_fact as it acts as a fact table containing all values. The final table has 118 rows and 23 columns as shown by the figures below.

```
df_population1['DISTRICT'].nunique()

df_fact = df_population1.merge(df_health,how='left',left_on='DISTRICT', right_on = 'DISTRICT/TEHSIL')
```

```
118
```

Fig 5.6 – Final dataframe is created after merging with health data

```
df_fact
df_fact.dtypes
```

	DISTRICT	AREA (SQ. KM.)	TOTAL POPULATION	POPULATION DENSITY PER SQ. KM.	AVERAGE HOUSEHOLD SIZE	POPULATION 1998	1998-2017 AVERAGE ANNUAL GROWTH RATE	URBAN POPULATION	RURAL POPULATION	WORKED (INCLUDED UN PAID FAMILY WORKER)	...	URBAN HOUSING UNITS	R HOI
0	BANNU	1227	1167071	951.16	9.34	675667	2.91	49948	1117123	153252	...	5994	1
1	LAKKI MARWAT	3164	875744	276.78	8.71	490025	3.10	89252	786492	122182	...	10922	
2	DERA ISMAIL KHAN	7326	1625088	221.82	7.96	852995	3.44	360218	1264870	314317	...	49896	1
3	KOHISTAN	7492	784711	104.74	7.84	472570	2.70	0	784711	132638	...	0	
4	TORGHAR	454	171349	377.42	6.49	174682	-0.10	0	171349	25347	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
113	LORALAI	8018	397423	49.57	7.00	250147	2.46	64891	332532	54039	...	8548	
114	MUSAKHEL	5728	167243	29.20	6.78	134056	1.17	14135	153108	19935	...	2076	
115	SHERANI	4310	152952	35.49	7.30	81684	3.35	0	152952	9477	...	0	
116	ZHOB	15987	310354	19.41	6.63	193458	2.51	46164	264190	33745	...	6614	
117	ISLAMABAD	906	2003368	2211.22	5.86	805235	4.90	1009003	994365	486288	...	167695	1

118 rows × 23 columns

Fig 5.6 – Final dataframe (df_fact)

In addition, 5 more columns were created and added to the final dataframe to analyse the results in a broader perspective. Such columns include urban, rural and total housing density with two more columns (Urban population density and rural population density).

```
df_fact['URBAN POPULATION DENSITY'] = df_fact['URBAN POPULATION']/df_fact['AREA (SQ. KM.)']
df_fact['RURAL POPULATION DENSITY'] = df_fact['RURAL POPULATION']/df_fact['AREA (SQ. KM.)']
df_fact['TOTAL HOUSING DENSITY'] = df_fact['OVERALL HOUSING UNITS']/df_fact['AREA (SQ. KM.)']
df_fact['URBAN HOUSING DENSITY'] = df_fact['URBAN HOUSING UNITS']/df_fact['AREA (SQ. KM.)']
df_fact['RURAL HOUSING DENSITY'] = df_fact['RURAL HOUSING UNITS']/df_fact['AREA (SQ. KM.)']
```

```
df_fact.to_csv('FinalFactTable.csv')
```

Fig 5.7 – computing and adding the new columns

```
DISTRICT                                    object
AREA (SQ. KM.)                               int64
TOTAL POPULATION                             int64
POPULATION DENSITY PER SQ. KM.             float64
AVERAGE HOUSEHOLD SIZE                     float64
POPULATION 1998                              int64
1998-2017 AVERAGE ANNUAL GROWTH RATE       float64
URBAN POPULATION                             int64
RURAL POPULATION                             int64
WORKED (INCLUDED UN PAID FAMILY WORKER)      int64
SEEKING WORK                                 int64
STUDENT                                      int64
OVERALL HOUSING UNITS                        int64
URBAN HOUSING UNITS                          int64
RURAL HOUSING UNITS                          int64
BELOW PRIMARY                                int64
PRIMARY                                      int64
MIDDLE                                       int64
MATRIC                                       int64
INTERMEDIATE                                 int64
GRADUATE                                     int64
DISTRICT/TEHSIL                             object
MEDICAL FACILITIES                         float64
URBAN POPULATION DENSITY                   float64
RURAL POPULATION DENSITY                   float64
TOTAL HOUSING DENSITY                      float64
URBAN HOUSING DENSITY                      float64
RURAL HOUSING DENSITY                      float64
dtype: object
```

Fig 5.8 – Datatypes of the final dataframe

Next up comes the analysis part where correlation is the first thing to investigate is the correlation between these variables (Zhang, McDonnell, Zadok, & Mueller, 2014) . This initial test of correlation will help select the variables needed for initial development of machine learning model. Figure 5.9 shows the use of libraries such as seaborn and matplotlib to plot a correlation heatmap.

```
import seaborn as sb
import matplotlib.pyplot as mp
```

```
mp.rcParams.update({'font.size': 15})
fig, ax = mp.subplots(figsize=(30,25))
dataplot=sb.heatmap(df_fact.corr(), ax=ax)
mp.show()
```

Fig 5.9 – generating correlation heatmap using seaborn

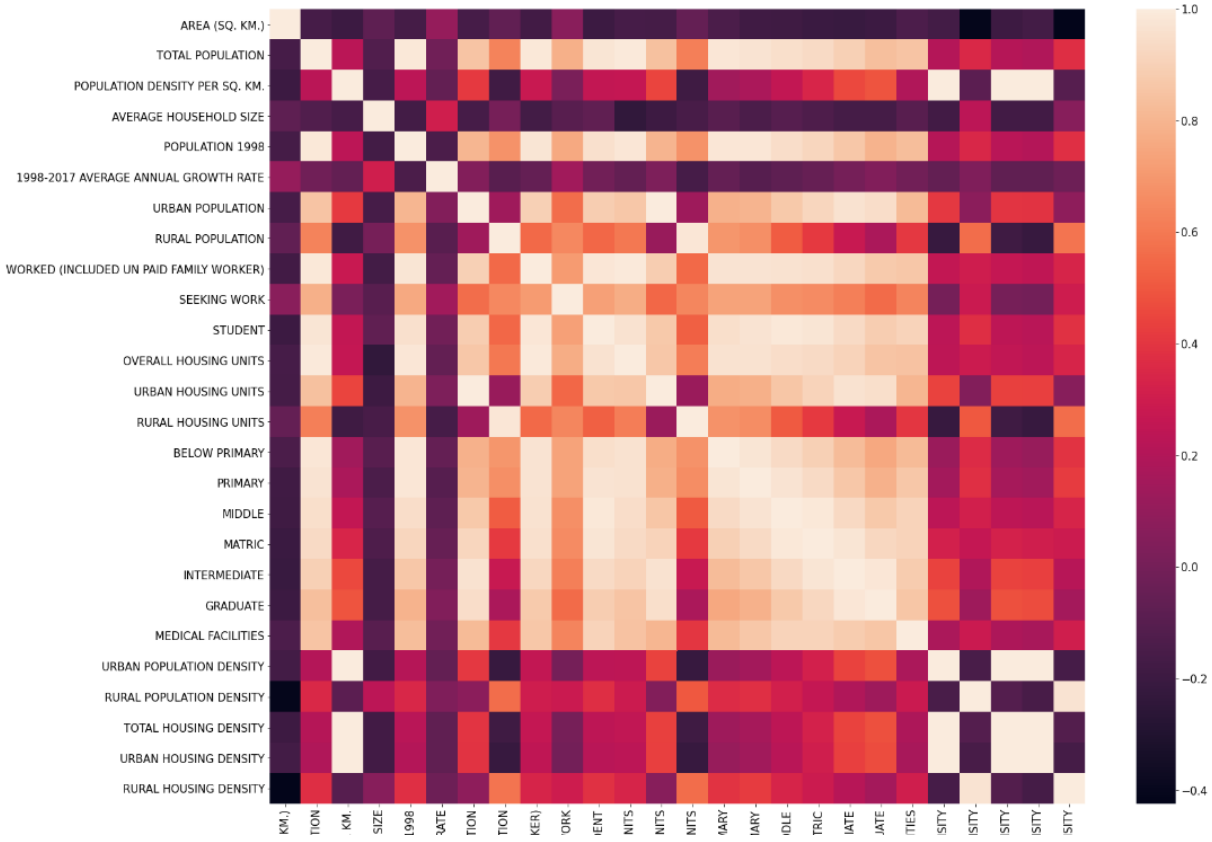


Fig 5.10 – Correlation heatmap with labels of values.

Fig 5.10 shows the correlation between the variables using a heatmap. It is quite evident from the heatmap that there is high amount of correlation between the variables. To get more detailed information, it is better to display numerical values of these correlations. The following figure (fig 5.11) shows the correlation heatmap with values displayed to analyse the correlations in a deep manner.

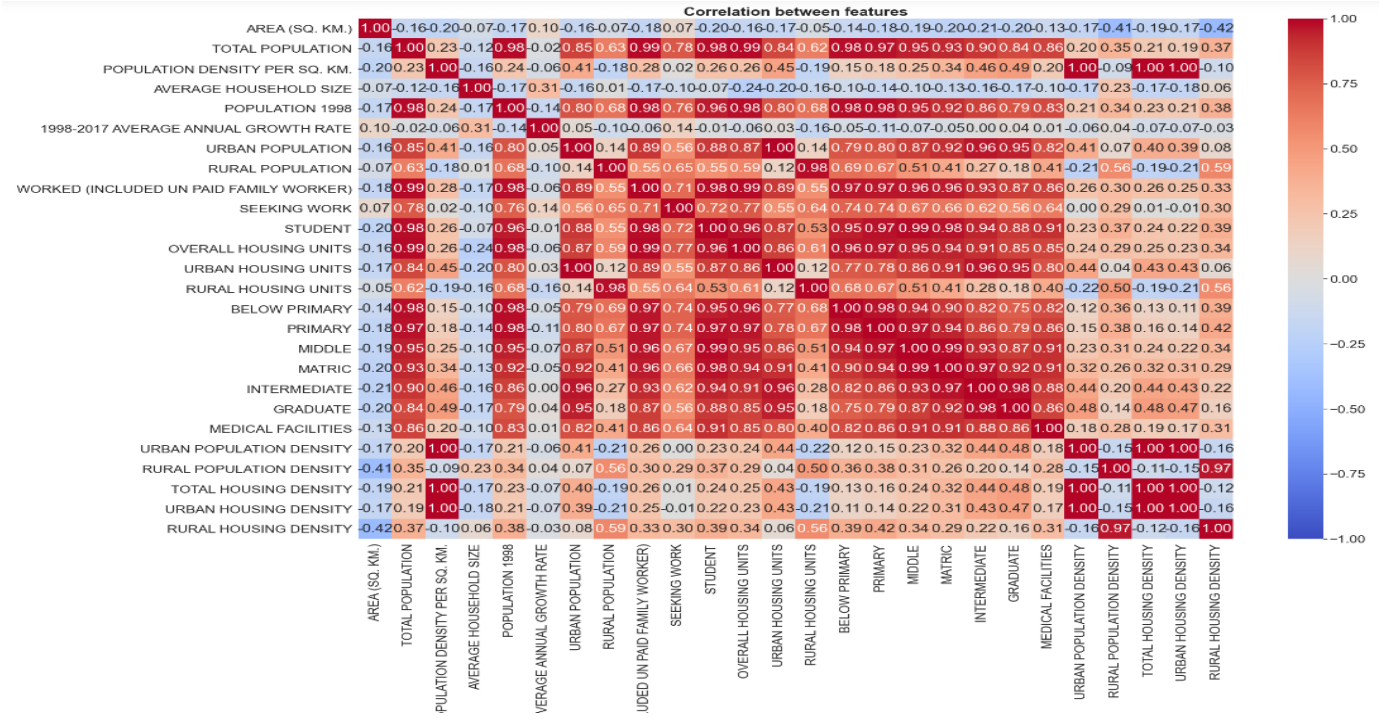


Fig 5.11 – Correlation values and heatmap

The high correlation between the variables can be witnessed because these variables are linked to each other directly or indirectly. Since there are so many columns being correlated, it is better to use intuition and choose only a few columns for the following analysis.
Once the columns are selected for final analysis, pair-plots will be visualised to check if there is any linear relationship between the variables and if there are any outliers in the data (Sahoo, Samal, Pramanik, & Pani, 2019). Figure 5.12 shows the pair plots generated using the seaborn library.

5.2 Variable selection and Model generation

```
sb.set(font_scale=1.3)
sb.pairplot(data=df_fact_selectfeature)
```

```
<seaborn.axisgrid.PairGrid at 0x228d775b310>
```

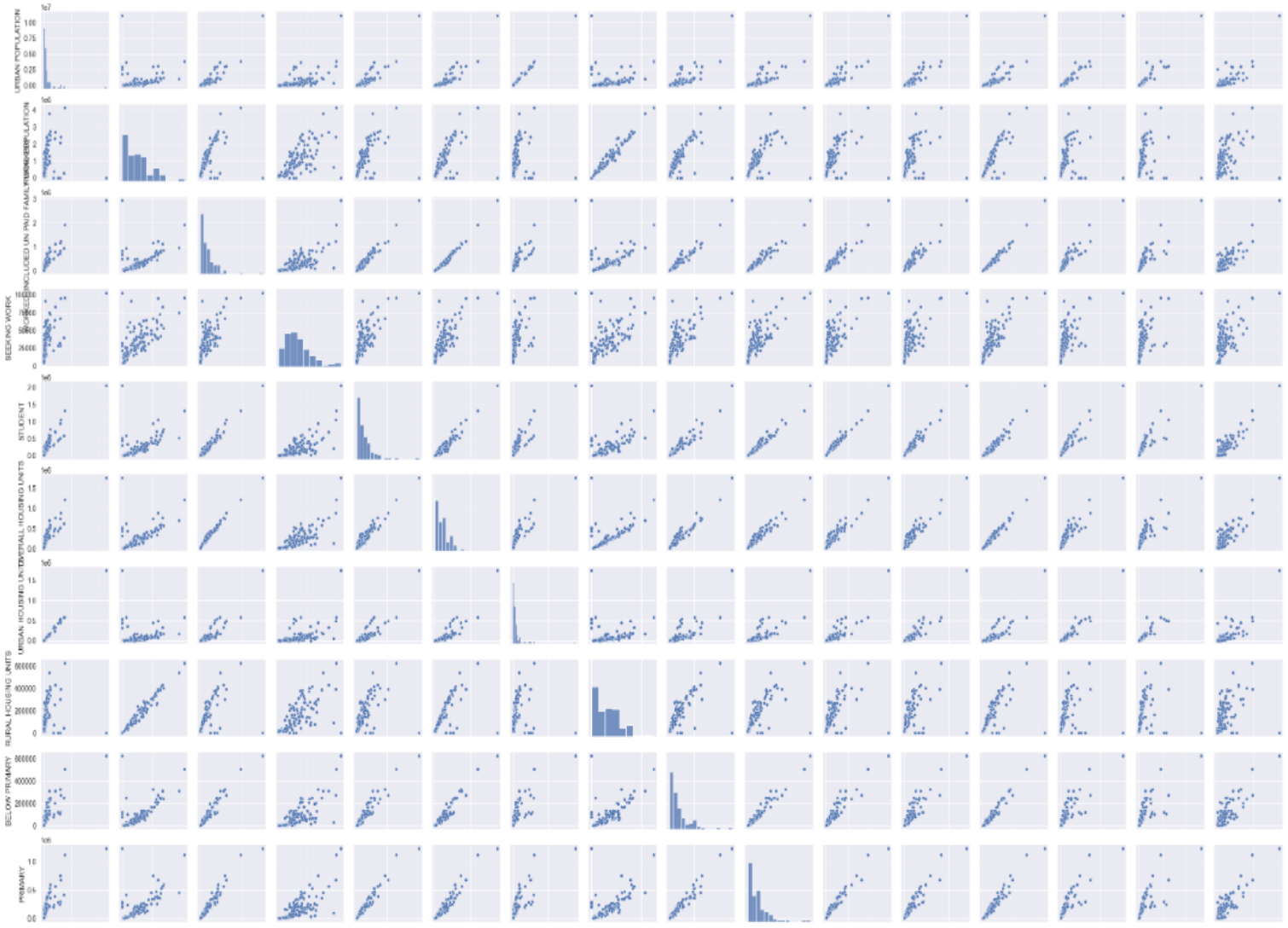


Fig 5.12 – Pair plots generated using seaborn to investigate linear relationship

In figure 5.12, it is quite evident that the variables have a linear relationship but there are still some outliers in the plots which need to be dealt with before creating the machine learning

model. There are number of ways to deal with outliers depending on the situation at hand but to eliminate the outliers, it is mandatory to understand as to why they are present in the data (Kwak & Kim, 2017). In current situation, there are outliers in the plotted pair-plots because these outliers are cities with high development which has extremely high numbers for the selected variables. For instance, Islamabad, Lahore, and Karachi are totally urbanized and are amongst the most populous districts (Khan & Adeel, 2017). The display of these districts as outliers in the above figure is evidence to the fact that population density in Pakistan is unfairly distributed along with other factors. Analysing as to why these outliers are present will help reach the research goals and objectives.

```
for i in dfdd.keys():
    if (i != 'DISTRICT'):
        q_low = dfdd[i].quantile(0.01)
        q_hi  = dfdd[i].quantile(0.99)
        median = dfdd.loc[(dfdd[i] < q_hi) & (dfdd[i] > q_low), i].median()
        dfdd.loc[(dfdd[i] > q_hi) |(dfdd[i] < q_low), i] = np.nan
        dfdd.fillna(median, inplace = True)
        dfdd.loc[(dfdd[i] > q_hi) |(dfdd[i] < q_low), i] = np.nan
```

Fig 5.13 – Dealing with outliers by replacing with median value

The figure above (Fig 5.13) shows the mechanism to deal with the outliers. The values of outliers have been replaced with the median value because the mean values are highly influenced by the outliers.

After replacing the outliers with median values, the pair plotting is conducted again to observe any differences. The following figure (fig 5.14) shows the pair plot after the outliers are dealt with by replacing them with median values.

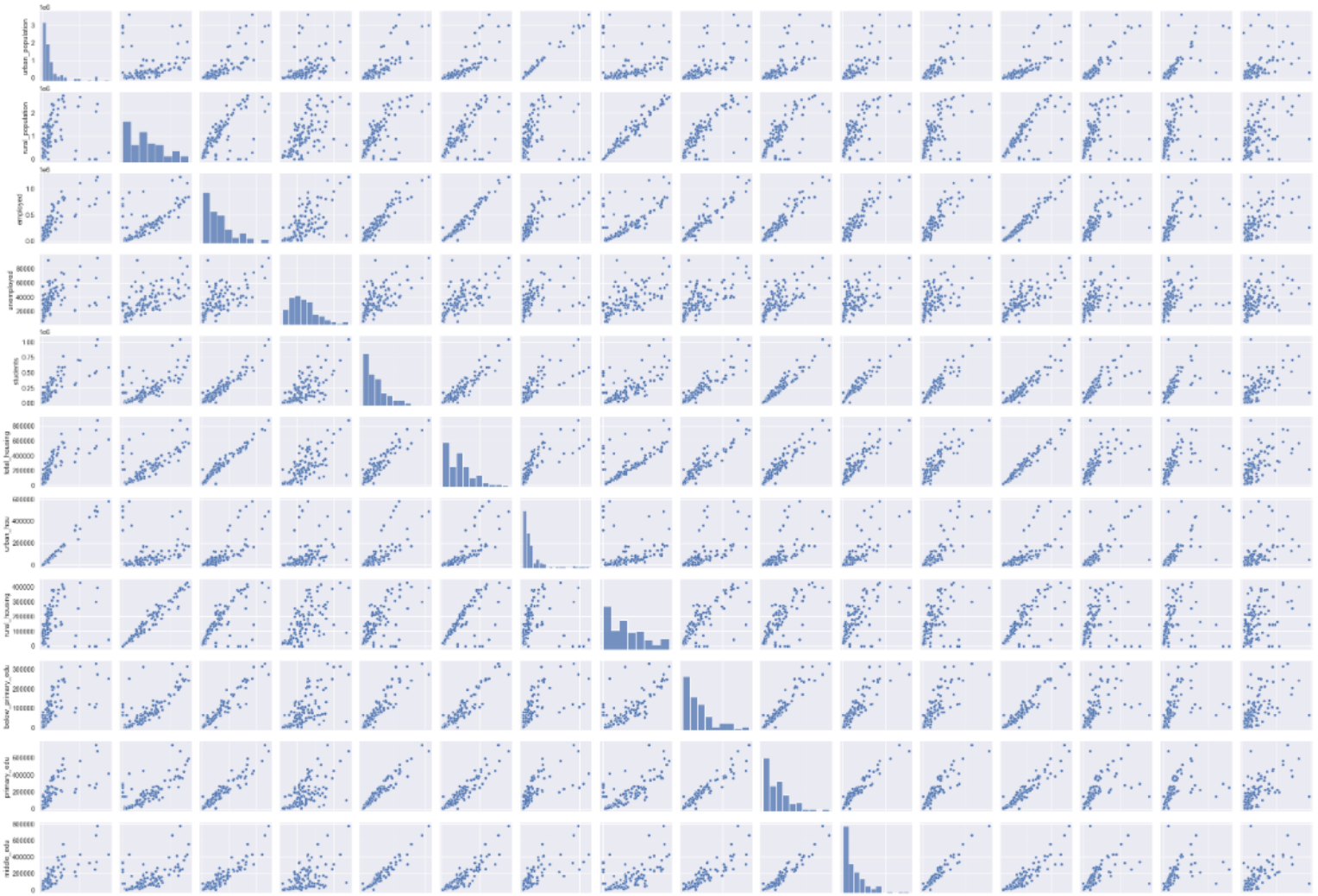


Fig 5.14 – Pair-plots after dealing with outliers

The mentioned pair-plots are now clear of any outliers, and it is apparent that the variables have a linear relationship. Digging deeper into the investigation of pair-plots, an imaginary line can also be witnessed in the graphs which means that linear regression will be suitable to use to create the research model (Andersen & Bro, 2010).

Moreover, the normality check of data is necessary before proceeding towards creation of the research model. There are a few tests to check the normality of the data such as skewness test. In the following research, a skewness test is performed for before and after the outliers are removed. Figure 5.15 shows the skewness values before and after the outliers are dealt with.

Skewness before after

```
for i in df_dd.keys():
    if i != 'DISTRICT':
        print(i, scipy.stats.skew(df_dd[i], axis = 0, bias = True))
print("df_dd", df_dd.shape[0])
print('-----------------------------------------------------------')
for i in dfdd.keys():
    if i != 'DISTRICT':
        print(i, scipy.stats.skew(dfdd[i], axis = 0, bias = True))
print("dfdd", dfdd.shape[0])
```

```
urban_population 5.65809680114633
rural_population 1.0613448821650024
employed 3.1834762692543
unemployed 1.0359513525203223
students 3.1014181616543732
total_housing 2.5987831369451424
urban_housing 5.441863349813389
rural_housing 0.9713028323751104
below_primary_edu 2.3466103110300653
primary_edu 2.4356453835108933
middle_edu 3.1885860205953622
matric_edu 3.5591644994969824
total_population 2.6404581866442918
intermediate_edu 4.257692491875319
graduate 4.780360305315624
medical_facilities 3.1505446875500724
df_dd 118
-----------------------------------------------------------
urban_population 2.490699095514837
rural_population 0.6651969060270649
employed 1.199613104471171
unemployed 0.9012564799145713
students 1.4851119139239062
total_housing 1.0242372497174133
urban_housing 2.5538089981036487
rural_housing 0.6374873327556884
below_primary_edu 1.2892353495140225
```

Fig 5.15 – Skewness values before and after replacing outlier values

Quite evident from the fig 5.15 is how the values have come closer to 0 after replacing the outlier values. Most of the variables in the research are positively skewed as shown by the data. Next up is determining the R-Squared value with respect to the target variable (total_population) (Kasuya, 2018). The following figure (fig 5.16) shows the r-squared values for independent variables.

```
y = dfdd['total_population']
X = dfdd.drop(['total_population', 'DISTRICT'], axis = 1)
function_dict = {'predictor': [], 'r-squared':[]}
for col in X.columns:
    selected_X = X[[col]]
    model = sm.OLS(y, sm.add_constant(selected_X)).fit()
    y_preds = model.predict(sm.add_constant(selected_X))
    function_dict['predictor'].append(col)
    r2 = np.corrcoef(y, y_preds)[0, 1]**2
    function_dict['r-squared'].append(r2)
function_df = pd.DataFrame(function_dict).sort_values(by=['r-squared'], ascending = False)
display(function_df)
```

	predictor	r-squared
5	total_housing	0.961935
2	employed	0.933687
8	below_primary_edu	0.907635
4	students	0.904915
9	primary_edu	0.903606
10	middle_edu	0.829273
11	matric_edu	0.756498
0	urban_population	0.599580
1	rural_population	0.566789
6	urban_hou	0.548664
7	rural_housing	0.545363
3	unemployed	0.523967
14	medical_facilities	0.471636
12	intermediate_edu	0.444884
13	graduate	0.348657

Fig 5.16 – R-Squared Values

The figure shows that total housing has the highest r-squared values. Hence, it will be taken along to take another r-squared test to see the new values along with the total_housing variable.

The following figure (fig 5.17) shows the r-squared values when total_housing is taken along as well.

R2 for all the predictor along with total_housing

```
selected_features = ['total_housing']
features_to_ignore = []
next_possible_feature (X_npf=X, y_npf=y, current_features=selected_features)
```

	predictor	r-squared
7	below_primary_edu	0.976321
4	students	0.975211
8	primary_edu	0.968802
1	rural_population	0.968585
2	employed	0.966463
9	middle_edu	0.966121
3	unemployed	0.963730
5	urban_hou	0.963719
13	medical_facilities	0.963251
6	rural_housing	0.962566
10	matric_edu	0.962379
12	graduate	0.962172
0	urban_population	0.962155
11	intermediate_edu	0.961939

Fig 5.17 – R-Squared values along with total_housing

Since the r-squared values for all variables displayed in figure 5.17 is very high, it would be effective to drop down more irrelevant and unnecessary columns. Columns such as *'primary_edu'*, '*below_primary_edu*', *'matric_edu'*, '*intermediate_edu*', and '*middle_edu*' because all of them can be deemed under the *'students'* variable.

After the columns are dropped, they are used to calculate the Variance Inflation Factor values (ViF Values). The ViF values will explain the nature of the independent variables with the dependent variable (Miles, 2014). In a nutshell, ViF depicts how well the variable is explained by other independent variables. A ViF above 10 indicates a high correlation and would be a cause of concern for the research. The following figure (fig 5.18) shows the ViF values for the selected columns with respect to their influence on dependent variable.

```
vif_2 = pd.DataFrame()
X_2 = X[['employed', 'unemployed', 'students', 'rural_population', 'total_housing', 'rural_housing', 'medical_facilities', 'g
vif_2["features"] = X_2.columns
vif_2["VIF"] = [variance_inflation_factor(X_2.values, i) \
                for i in range(len(X_2.columns))]
vif_2
```

	features	VIF
0	employed	65.305220
1	unemployed	9.728174
2	students	46.443745
3	rural_population	120.237405
4	total_housing	94.572091
5	rural_housing	115.036182
6	medical_facilities	7.474063
7	graduate	5.536573
8	urban_population	215.601194
9	urban_hou	207.963201

Fig 5.18 – ViF values initially recorded after selecting columns

The following table shows that *'medical_facilities'*, *'unemployed'*, and *'graduate'* have the most adequate ViF values as they are less than 10 whereas the other variables have extremely high ViF values.

Using these ViF values a prototype machine learning model is created using these three variables to get a basic idea of how well the model will look like using just these three features. OLS (Ordinary Least Square) regression is the type of regression used to conduct the research as it is used for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (ref.). The following figure (fig 5.19) shows the summary of the machine learning model created using these three variables (*'medical_facilities'*, *'unemployed'*, and *'graduate'*).

Modal with features having low VIF value

In [61]:
```
final_features = ['unemployed', 'medical_facilities', 'graduate']
final_X = X[final_features]
final_model = sm.OLS(y, sm.add_constant(final_X)).fit()
final_model.summary()
```

Out[61]:

OLS Regression Results

Dep. Variable:	total_population	**R-squared:**	0.706
Model:	OLS	**Adj. R-squared:**	0.698
Method:	Least Squares	**F-statistic:**	91.12
Date:	Wed, 14 Sep 2022	**Prob (F-statistic):**	3.81e-30
Time:	12:59:26	**Log-Likelihood:**	-1743.9
No. Observations:	118	**AIC:**	3496.
Df Residuals:	114	**BIC:**	3507.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.293e+05	1.27e+05	-1.021	0.310	-3.8e+05	1.22e+05
unemployed	33.4460	3.752	8.914	0.000	26.013	40.879
medical_facilities	7705.6774	2312.455	3.332	0.001	3124.721	1.23e+04
graduate	6.0870	1.601	3.802	0.000	2.915	9.259

Omnibus:	21.091	**Durbin-Watson:**	1.416
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	67.136
Skew:	-0.528	**Prob(JB):**	2.64e-15
Kurtosis:	6.541	**Cond. No.**	1.47e+05

Fig 5.19 – Summary of OLS Regression results using variables

The above information in the figure shows that the R-Square value of the model is 0.706 which is a decent value, but it still does not consider the rest of the variables. To tackle this problem, Principal Component Analysis can be performed (Ringner, 2008). As mentioned in the literature of this research, Principal Component analysis is a technique used to reduce the dimensionality of a dataset, while preserving as much 'variability' (i.e., statistical information) as possible.

Using principal component analysis would investigate the variables and create principal components depending on circumstances. The principal components are variables with their valuable information clustered together in a way which reduces the dimensionality. The following figures (fig 5.20 a-d) explain the use of Principal Component Analysis in the conducted research.

Standard Scale before Dimensionality reduction (PCA)

```
features = ['urban_population', 'rural_population', 'employed', 'students', 'total_housing', 'urban_hou', 'rural_housing']
xx = dfdd.loc[:, features].values
xx = StandardScaler().fit_transform(xx)
target = dfdd['total_population'].values
```

```
pca = PCA().fit(xx)
plt.rcParams["figure.figsize"] = (12,6)
fig, ax = plt.subplots()
xi = np.arange(1, 8, step=1)
y = np.cumsum(pca.explained_variance_ratio_)
plt.ylim(0.0,1.1)
plt.plot(xi, y, marker='o', linestyle='--', color='b')
plt.xlabel('Number of Components')
plt.xticks(np.arange(0, 11, step=1)) #change from 0-based array index to 1-based human-readable label
plt.ylabel('Cumulative variance (%)')
plt.title('The number of components needed to explain variance')
plt.axhline(y=0.95, color='r', linestyle='-')
plt.text(0.5, 0.85, '95% cut-off threshold', color = 'red', fontsize=16)
ax.grid(axis='x')
plt.show()
```

fig 5.20a – Using features with high ViF values to create a Principal Component

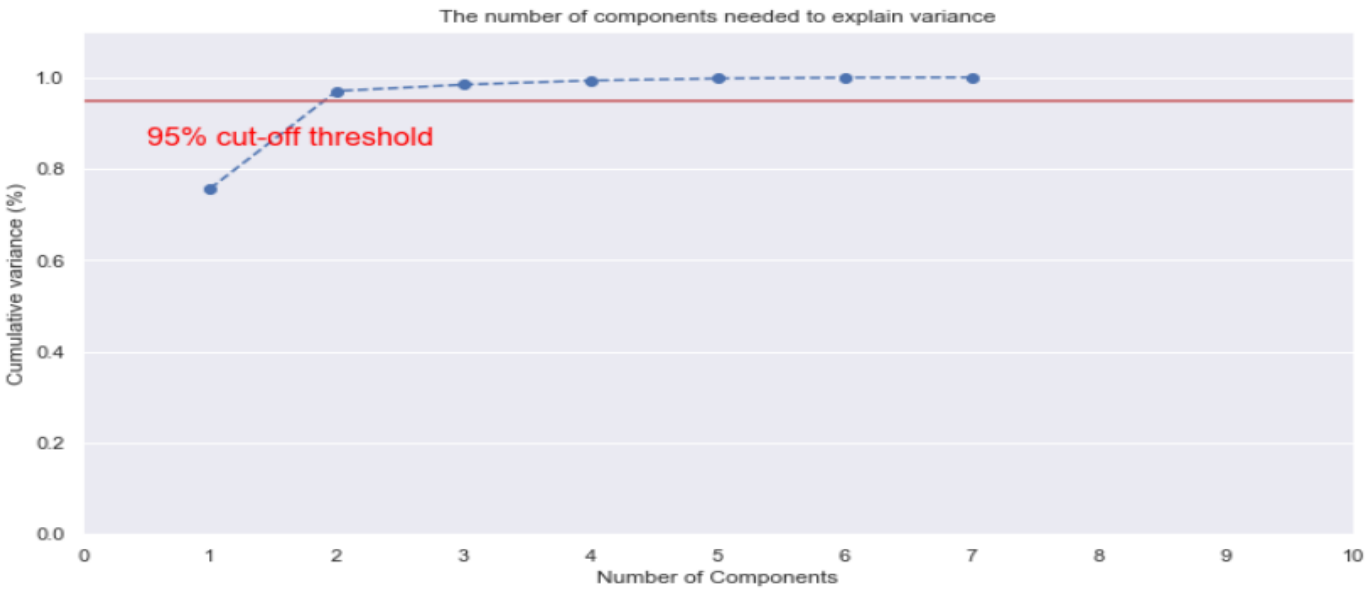


Fig 5.20b – No. of components needed to explain the variance

PCA final dataframe

```
pca =PCA(n_components=1)
pca.fit(xx)
principalComponents = pca.fit_transform(xx)
principalDf = pd.DataFrame(data = principalComponents
             , columns = ['pc1'])
targetDf = pd.DataFrame(data = target, columns = ['total_population'])
finalDf = pd.concat([principalDf, targetDf], axis = 1)
finalDf
```

55]: PCA(n_components=1)

55]:

	pc1	total_population
0	-1.149856	1167071.0
1	-1.527496	875744.0
2	-0.162902	1625088.0
3	-1.730632	784711.0
4	-2.569516	171349.0
...	...	...
113	-2.228960	397423.0
114	-2.598840	167243.0
115	-2.255780	152952.0
116	-2.370858	310354.0
117	1.427461	2003368.0

118 rows × 2 columns

Fig 5.20c – Final dataframe after Principal Component Analysis

PCA with other features which have low VIF value

```
finalDff = pd.concat([final_X, finalDf], axis = 1)
finalDff
```

6]:

	unemployed	medical_facilities	graduate	pc1	total_population
0	30973.0	20.0	21722.0	-1.149856	1167071.0
1	21728.0	15.0	13017.0	-1.527496	875744.0
2	25058.0	24.0	29365.0	-0.162902	1625088.0
3	55409.0	6.0	3163.0	-1.730632	784711.0
4	30816.5	24.5	22410.5	-2.569516	171349.0
...	...	...	...	...	...
113	16852.0	40.0	5988.0	-2.228960	397423.0
114	7964.0	15.0	1526.0	-2.598840	167243.0
115	6439.0	6.0	22410.5	-2.255780	152952.0
116	14366.0	27.0	4820.0	-2.370858	310354.0
117	29663.0	101.0	165599.0	1.427461	2003368.0

118 rows × 5 columns

Fig 5.20d – Principal components and other variables with low ViF values in a final dataframe

The above illustrated graphs and tables explain how Principal components were created using principal components analysis. All the variables with high ViF values were used with total population as the target variable. This helps in obtaining a final dataframe which has those three variables with low ViF value and the principal component. There is only one principal component because 95% of the data can be explained using 1 component (refer to fig 5.20b). After the final dataframe is generated, OLS regression method is applied to create a machine learning model. The summary of the final model is provided in the figure below (fig 5.21).

```
smfOLS = sm.OLS(y_train, X_train).fit()
y_pred = smfOLS.predict(X_test)
smfOLS.summary()
```

]:

OLS Regression Results

Dep. Variable:	total_population	**R-squared (uncentered):**	0.948
Model:	OLS	**Adj. R-squared (uncentered):**	0.945
Method:	Least Squares	**F-statistic:**	355.9
Date:	Wed, 14 Sep 2022	**Prob (F-statistic):**	3.06e-49
Time:	13:00:42	**Log-Likelihood:**	-1179.3
No. Observations:	82	**AIC:**	2367.
Df Residuals:	78	**BIC:**	2376.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
unemployed	35.2335	2.112	16.683	0.000	31.029	39.438
medical_facilities	1874.7000	2007.758	0.934	0.353	-2122.439	5871.839
graduate	4.8901	1.772	2.759	0.007	1.362	8.419
pc1	2.245e+05	2.46e+04	9.123	0.000	1.75e+05	2.73e+05

Omnibus:	11.517	**Durbin-Watson:**	1.785
Prob(Omnibus):	0.003	**Jarque-Bera (JB):**	11.804
Skew:	-0.848	**Prob(JB):**	0.00273
Kurtosis:	3.762	**Cond. No.**	3.04e+04

Fig 5.21 – Summary of final model with key variables and principal components

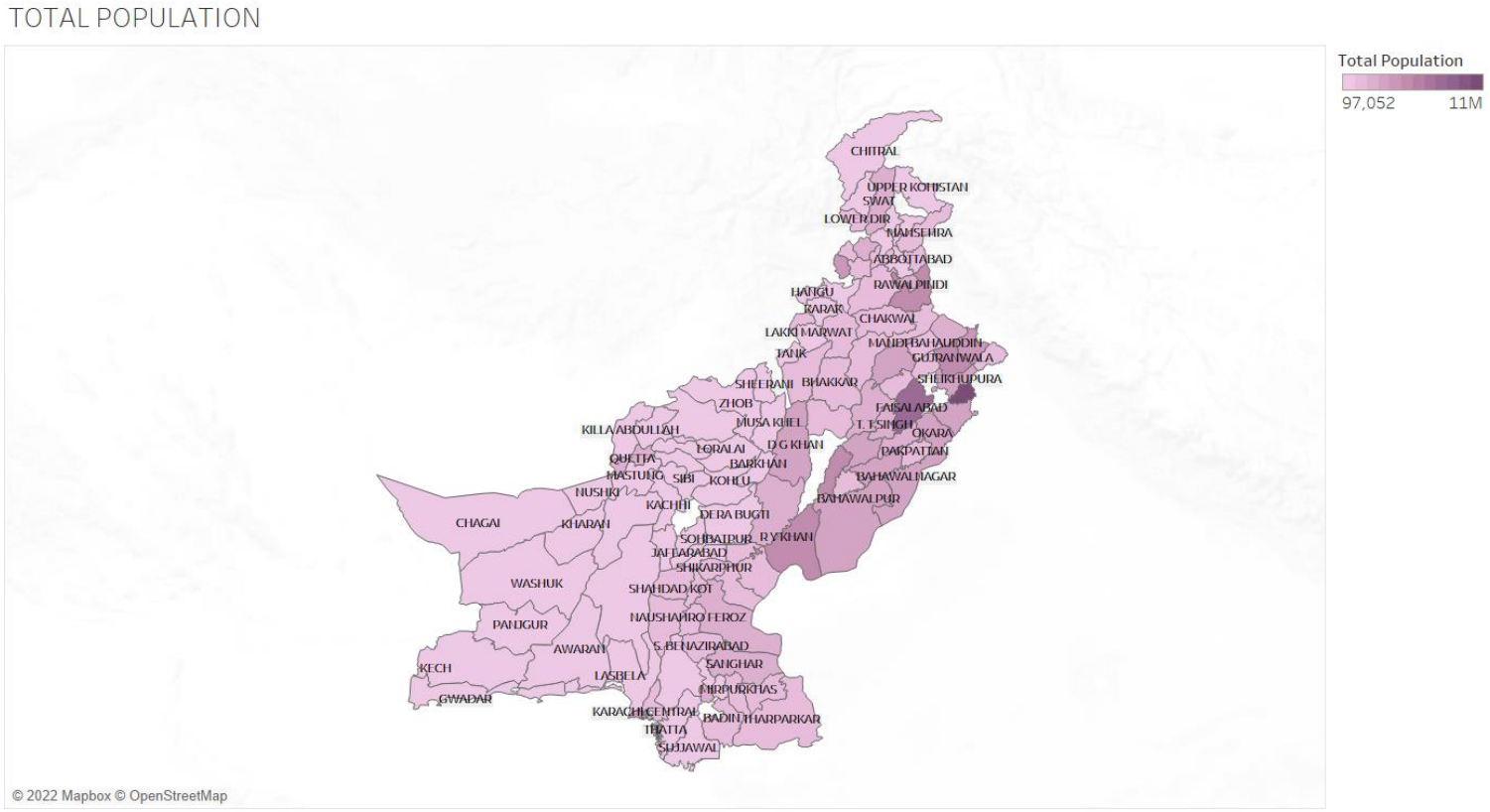
The final summary of the model shows that it has 0.948 R-Squared value which serves as a good value. It indicates a high value of variation of dependent variable explained by independent variables. The model summary also indicates that the coefficient values is least for graduation and highest for medical facilities when it comes to factors excluding the principal component. Considering this it can be concluded that this final model is valid and accurate using the following variables with the addition of the principal component.

6. Insights and Findings

Since the model has been created, it is essential to visualize the variables in a more descriptive approach. Visualisations using shapefiles to generate spatial visualisations is one way of exploring the data of Pakistan. The dataset collected from Pakistan Bureau of Statistics already had a shapefiles folder which contained shapefiles for administrative boundaries. Since the current research is conducted based on Districts, it is better to use the district boundaries shapefile to explore the data on a map using Tableau.

The shapefiles were joined with the final fact table to visualize the data along with spatial data provided in shapefiles. The following figures (fig 6.2 - 6.7) show the key variables and their distribution over the map of Pakistan (fig 6.1). The maps are colored according to the heat of data like a heatmap.

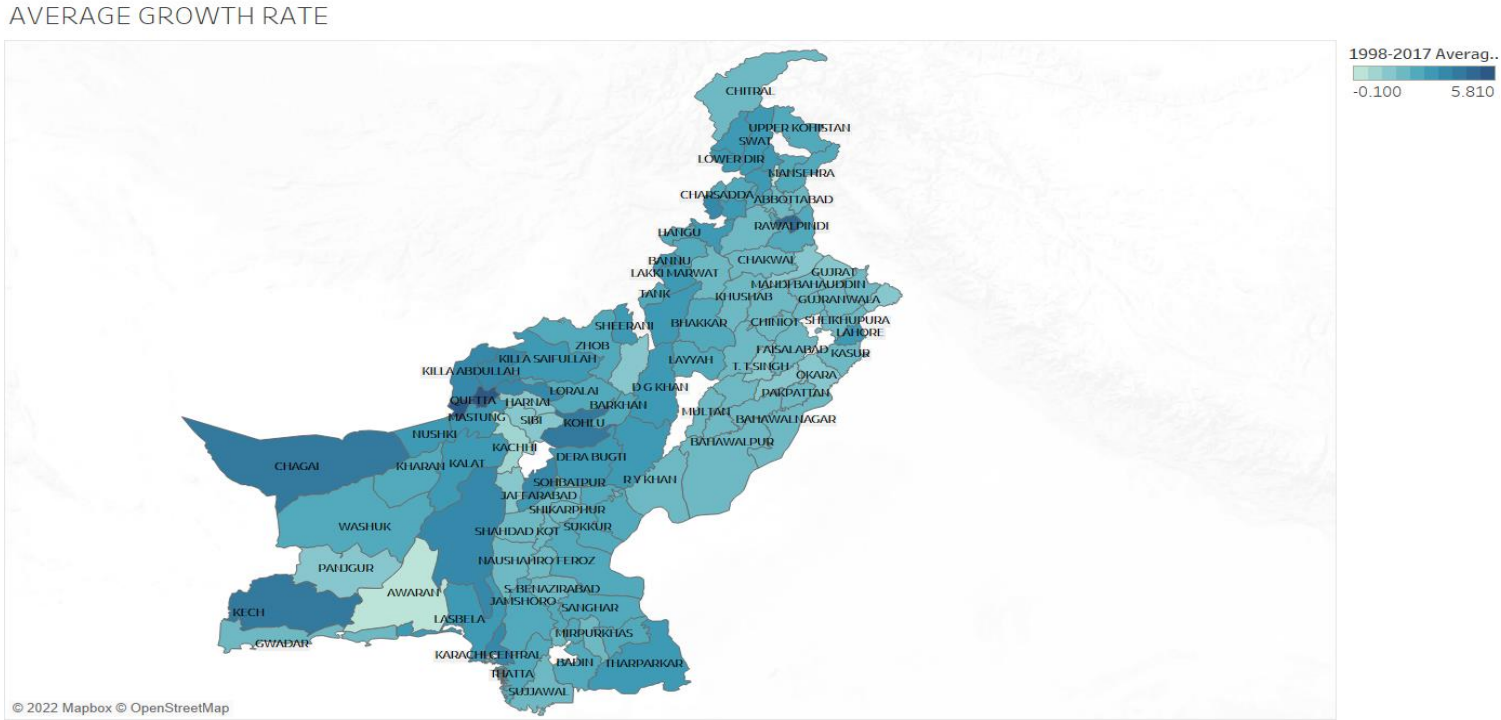
Fig 6.1 Map of Pakistan with labelled provinces and key rivers



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Total Population. The marks are labeled by District. Details are shown for District.

Fig 6.2 – Map of Pakistan showing distribution of Total Population

In fig 6.2 it is quite evident that most of the country's population lies in the Punjab and Sindh province with some exceptions in Khyber Pakhtunkhwa. Lahore, Karachi, and Islamabad Districts seem to be one of the most populated districts. Punjab is clearly the most populated province with the greatest number of populated districts.



Map based on Longitude (generated) and Latitude (generated). Color shows sum of 1998-2017 Average Annual Growth Rate. The marks are labeled by District. Details are shown for District.

Fig 6.3 – Map of Pakistan showing Average Growth of Population from 1998 to 2017

The figure above (fig 6.3) explains the distribution of average growth of population in Pakistan from 1998 to 2017. Balochistan is the province which has the greatest number of districts with highest average population growth in comparison to other provinces like Punjab, KPK and Sindh. Quetta and Chagai are one of the most fastest growing districts in Balochistan.

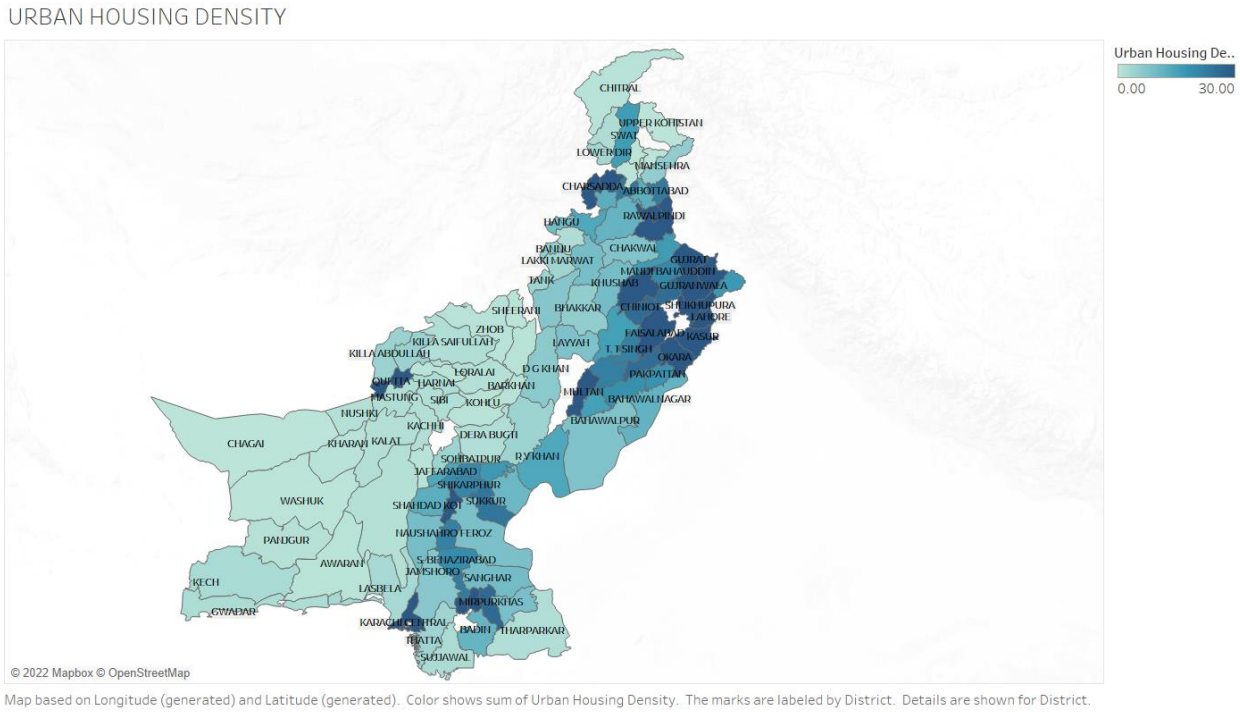


Fig 6.4 – Map of Pakistan showing the distribution of Urban Housing density

The urban housing density distribution over Pakistan can be interpreted using the figure (fig 6.4). The figure gives an impression of high urban housing density in Punjab and some areas of Sindh and KPK, but Punjab has the highest number of districts with high urban housing density. Districts around Lahore and Islamabad tend to be the densest districts in terms of Urban Housing.

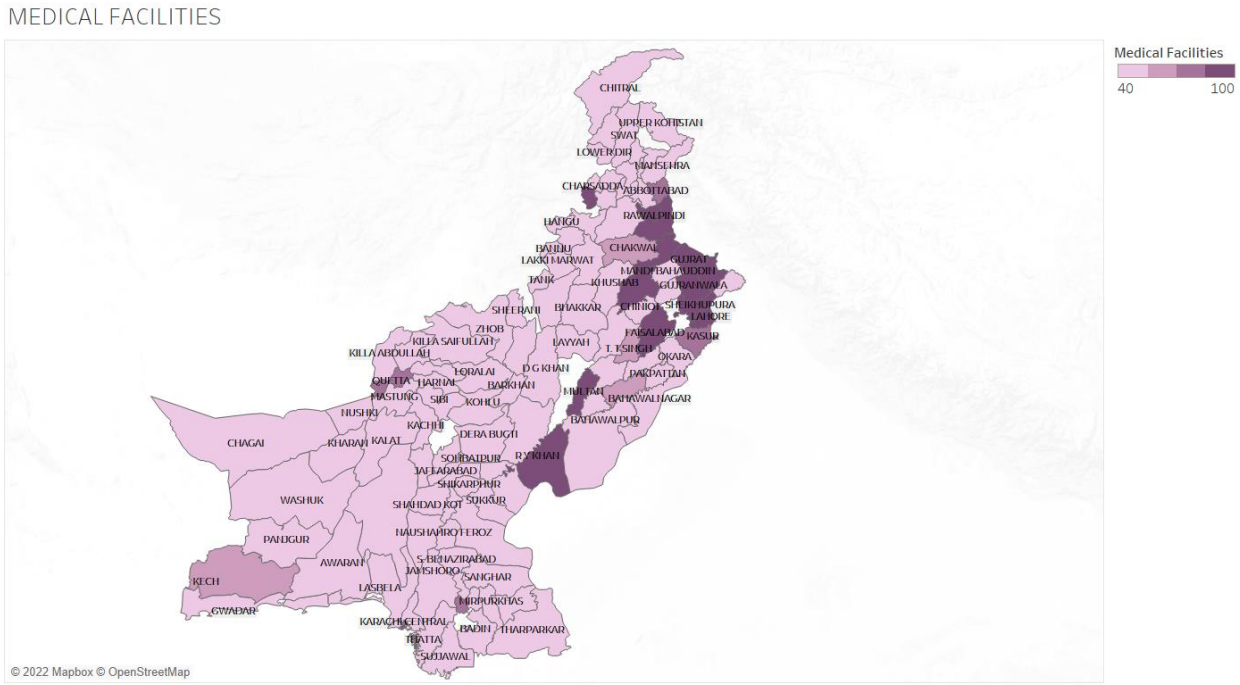


Fig 6.5 – Map of Pakistan Showing Medical Facilities

The figure for visualizing medical data on Pakistan map shows that the data is highly diverse. Districts in Punjab which are located near Islamabad and Lahore have high number of medical facilities whereas Balochistan hardly has any District with high number of medical facilities. The districts in Karachi division are the districts with highest number of medical facilities in Sindh.

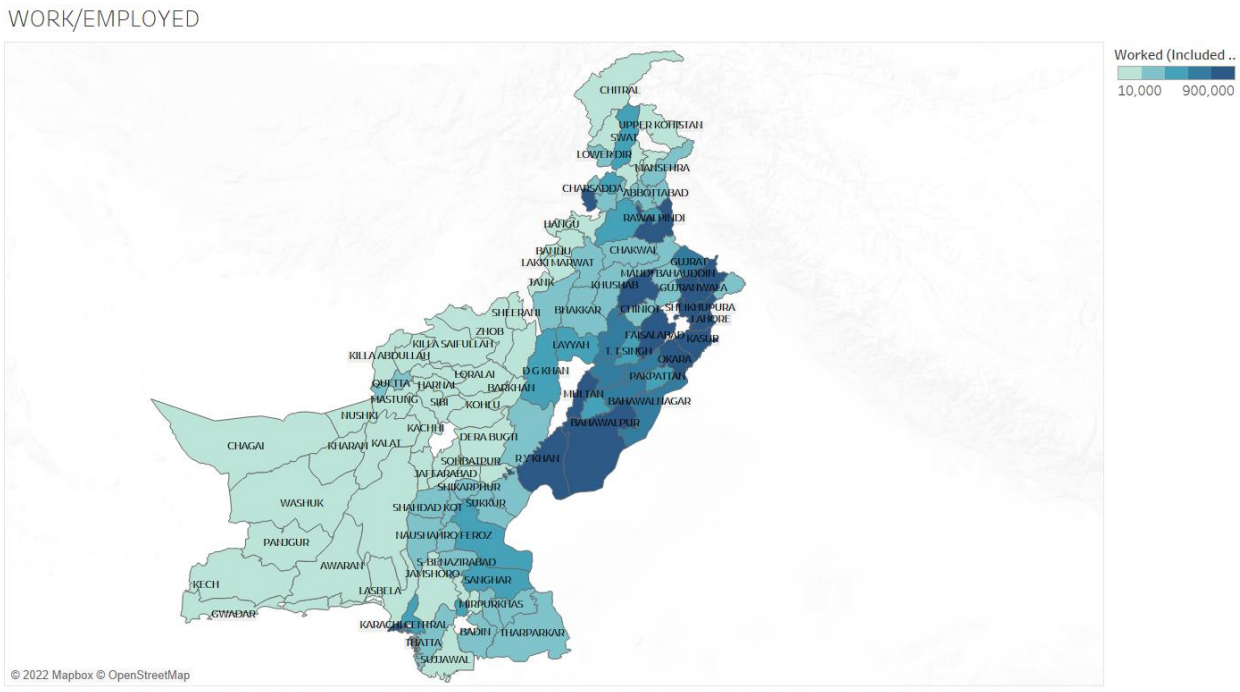
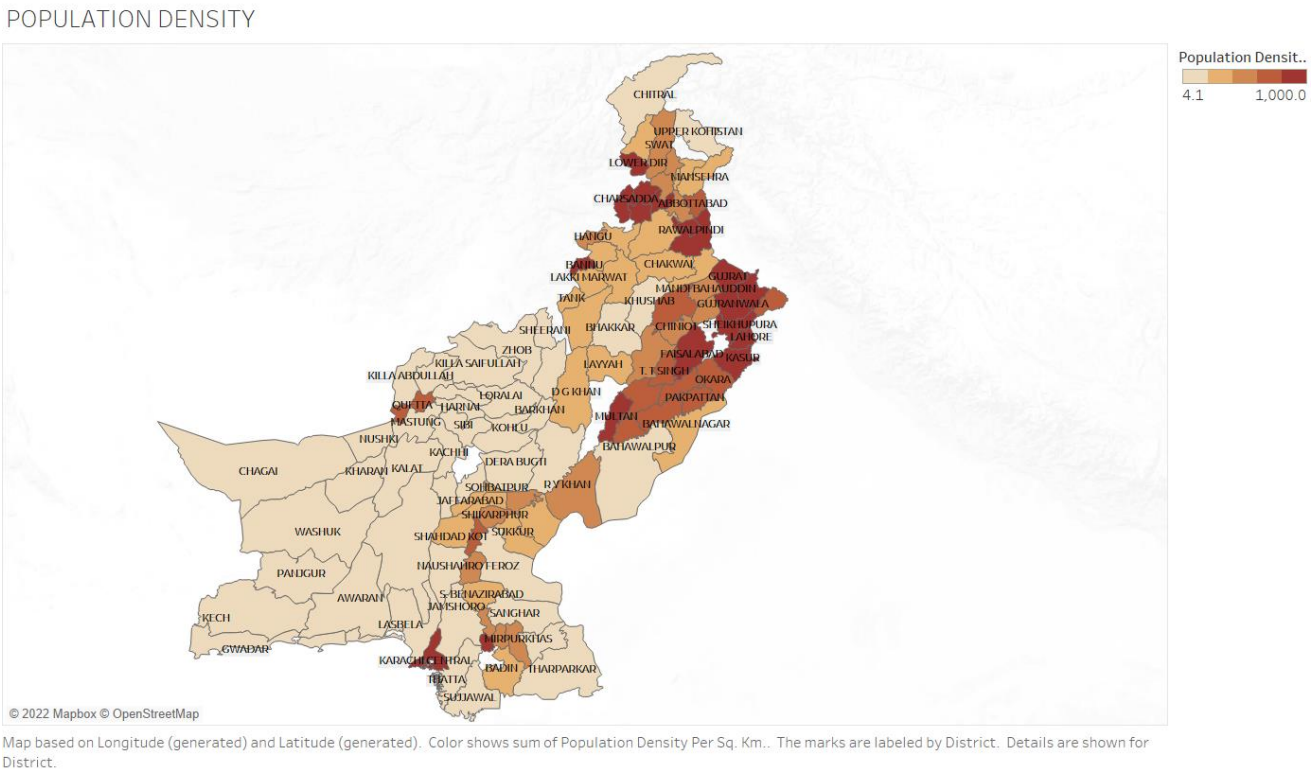


Fig 6.6 – Map of Pakistan Showing employed people

Moving on, the data for employed personnel is displayed on the map of Pakistan to evaluate the performance of districts in terms of employment. Visible from the fig 6.6, Punjab has the highest employment as compared to other provinces. In Sindh, districts of Karachi division have the highest employment. Balochistan has the lowest score in terms of employment.



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Population Density Per Sq. Km.. The marks are labeled by District. Details are shown for District.

Fig 6.7 – Map of Pakistan Showing population density

The final figure (fig 6.7) shows the population density of Pakistan spread across the provinces. Districts in Punjab, Sindh, and some areas of KPK have the highest population density with Punjab topping the chart. This shows that Punjab is the most densely populated province of Pakistan.

7. Research Conclusions

7.1 Research Discussion

The main deliverable of this research was to identify variables that are the prime reasons for rapid urbanization and population in Pakistan. Completing the literature review helped in eliminating all doubts regarding the research by setting a clear approach as to what to do. The research aims were to be achieved by completing the research objectives:

- Complete a literature review using past research papers, articles, etc. to understand current research
- Gather data from concerned organisations regarding the given research
- Identify key variables and factors in the dataset to be used in further research

- Create statistical models using Machine Learning
- Measure accuracy of the model to get valid results
- Test the system for anomalies
- Draw visualizations to answer relevant questions for the research
- Evaluate results and draw future scope for the research

The main aim of the research was to detect key variables by using a machine learning model. These variables would constitute the most towards the issue at hand. Since the model is considered valid and accurate for the research, three variables are said to be the most influential variables for the current research issue. Unemployment, Medical Facilities and Graduate variables are the most highly influential variables. Unemployment and graduates can be termed together as well for a broader perspective. For example, Unemployed people would also be looking for work opportunities in a city same as graduates. Medical facilities are another feature that is highly influential for the conducted research. Hence, according to the model, districts with large number of job opportunities and medical facilities would have the highest number of urbanization and population in it. This is another reason as to why in recent time the Pakistani Government initiated the Naya Pakistan Housing Scheme to tackle the housing and urbanization crisis in Pakistan (Khalil & Nadeem, 2019). It can also be observed that most of the developed districts in Pakistan are positioned near Indus River (refer to Fig 6.1 for Pakistan map with rivers) which gives a better future scope to the following research in terms of developing other districts.

7.2 Limitations and Difficulties during research

The health data used in the research was collected in 2012 whereas in 2013 some new districts were created from the previous ones. However, the census data was collected in 2017 and would have a few new districts as compared to our health dataset. To make sure the data is valid, certain checks and cleaning of data was required. To make sure the research is un-biased and fair, it was important to manually check and correctly add these new districts with their respective Tehsils. The data for the new districts and their tehsils must be dropped from previous data and joined together to make sure we do not get or add repeated values for our research.

During the initial stages of the data cleaning stage, it was fair to drop Federal Administered Tribal Areas, FR (Federal Regions and Agencies) as they were given the provincial autonomous powers in 2018 by merging them with the KPK province (previously NWFP). Hence, the autonomous control to such areas came after our data was collected it is not fair to include such areas in our analysis.

The housing and population dataset was used to create additional columns that would indicate powerful metrics that would enable the research to be more prolific. Such metrics include Rural

Housing Density, Urban Housing Density, Overall Housing density, Urban population density and Rural population density.

The densities were calculated as follows:

Population or Housing Units / Area Size (in km. sq)

In the education dataset, the Musakhel district had some invalid naming conventions that was later causing errors in merging our data frames to get correct values. It was essential to manually check and correct the values for Musakhel District in Education Sector. The naming convention issue was constant in the data cleaning stage as all districts are named in Urdu and when written in English would have different spellings. Manually checking and renaming the rows was required before our data could be merged into a final dataset (Rahm & Hai Do, 2000).

In the health dataset, the data is filtered to show only medical facilities in these districts because when analysing the situation at district level it is better to discuss standard medical facilities in the district such as Hospitals and treatment centres rather than dispensaries or pharmacies.

Some district had NaN values in urban housing and medical facilities because there are some districts in Pakistan where there is no urban area, and some do not even have a medical facility (Torghar District).

Furthermore, there were many districts that were created after the data was collected by the sources. It was not possible to add these districts in our analysis as the data timeline for the used datasets does not match with the timeline of these newly created districts (Babakhel, 2018).

7.3 Future Scope

Having identified the key factors driving the high urbanization and population rate in Pakistan, it can be very beneficial if other districts are developed using these key variables. In recent times, Pakistan government has planned towards tackling the population crisis by introducing new housing schemes and development of smart cities. The conducted research can help explore variables discussed in development of smart cities in Pakistan in the future. It is also necessary to note that it would be better to develop districts which are closer to these well-developed districts because in this way the transportation and logistic issues can be overcome as there will be less expenditure as well.

References

Abdi, H., & J. Williams, L. (2010). Principal component analysis. *WIREs Computational Statistics*.

Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*.

Babakhel, M. A. (2018, October 23). *New Districts*. Retrieved from Dawn News: https://www.dawn.com/news/1440719

Bari, M. (2020, 10 05). *Is rapid urbanization making Pakistan's cities less livable?* Retrieved from DW: https://www.dw.com/en/is-rapid-urbanization-making-pakistans-cities-less-livable/a-55162735

Ben Braiek, H., & Khomh, F. (2020). On testing machine learning programs. *Journal of Systems and software*.

Chen, D. Y. (2017). *Pandas For Everyone.* Pearson Education.

Darlington, R. B., & Hayes, A. F. (2017). *Regression Analysis and Linear Models.* The Guilford Press.

Desboulets, L. (2018). A Review on Variable Selection in Regression Analysis. *Econometrics*.

Ezekiel, M., & Fox, K. A. (1959). *Methods of correlation and regression analysis: Linear and curvilinear.* John Wiley.

Famili, A., Shen, W.-M., Weber, R., & Simoudis, E. (1997). Data Preprocessing and Intelligent Data Analysis. *Intelligent Data Analysis (Vol. 1)*.

Gogtay, N., & Thatte, U. (2017). Principles of Correlation Analysis. *Journal of The Association of Physicians of India*.

Hashim, A. (2018, May 24). *Pakistan parliament passes landmark tribal areas reform*. Retrieved from Al-Jazeera: https://www.aljazeera.com/news/2018/5/24/pakistan-parliament-passes-landmark-tribal-areas-reform

Hutcheson, G. D., & Moutinho, L. A. (2011). *The SAGE Dictionary of Quantitative Management Research.* Sage Publications Ltd.

Joliffe T, I., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *The Royal Society Publishing*.

Kasuya, E. (2018). On the use of r and r squared in correlation and regression. *Ecological Research*.

Khalil, I., & Nadeem, U. (2019). Optimising the Naya Pakistan Housing Policy Opportunity. *Tabadlab*.

Khan, S., & Adeel, D. M. (2017, October 17). *The curious case of urban population in Pakistan*. Retrieved from London School Of Economics:

https://blogs.lse.ac.uk/southasia/2017/10/17/the-curious-case-of-urban-population-in-pakistan/

Kim, R. S., Aloe, A. M., & Becker, B. J. (2017). Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results. *Basic and Applied Social Psychology*.

Kirk, A. (2016). *Data Visualisation: A Handbook for Data Driven Design.* SAGE Publications Ltd.

Kugelman, M. (2013). Urbanisation in Pakistan: causes and consequences. *NOREF Norwegian Peacebuilding Resource Centre*.

Kuo, L., & Mallick, B. (1998). Variable Selection for Regression Models. *Sankhyā: The Indian Journal of Statistics*, 65-81.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*.

Lee, I., & Jae Shin, Y. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, 157-170.

Mansfield, E. R., & Helms, B. P. (1982). Detecting Multicollinearity. *The American Statistician*.

Miles, J. (2014). Tolerance and Variance Inflation Factor. *Encyclopedia of Statistics in Behavioral Science*.

Morrow, V., Boddy, J., & Lamb, R. (2014). The ethics of secondary data analysis: Learning from the experience of sharing qualitative data from young people and their families in an international study of childhood poverty. *NOVELLA Working Paper: Narrative Research in Action*.

Nogueira, S., Brown, G., & Sechidis, K. (2018). On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research 18*, 1-54.

Poole, M. A., & O'Farrell, P. N. (1971). The Assumptions of the Linear Regression Model. *Transactions of the Institute of British Geographers*, 145-158.

Raheem, M. A., Udoh, N. S., & Gbolahan, A. T. (2019). Choosing Appropriate Regression Model in the Presence of Multicolinearity. *Open Journal of Statistics*.

Rahm, E., & Hai Do, H. (2000). Data Cleaning: Problems and Current Approaches. *Data Engineering IEEE*, 3-14.

Ringner, M. (2008, March). *What is principal component analysis?* Retrieved from Nature: https://www.nature.com/articles/nbt0308-303

Rong, S., & Bao-wen, Z. (2018). The research of regression model in machine learning field. *MATEC Web Conf.*

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*.

Shaikh, H., & Nabi, I. (2017, January 16). *The six biggest challenges facing Pakistan's urban future*. Retrieved from International Growth Centre: https://www.theigc.org/blog/the-six-biggest-challenges-facing-pakistans-urban-future/

Stancin, I., & Jovic, A. (2019). *An overview and comparison of free Python libraries for data mining and big data analysis.* Opatija, Croatia: IEEE.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*.

T. Pohlmann, J., & W. Leitner, D. (2003). A comparison of ordinary least squares and logistic regression. *The Ohio Journal of Science*.

Wazir, M. A., & Goujon, A. (2019). *Assessing the 2017 Census of Pakistan Using Demographic Analysis: A Sub-National Perspective .* Vienna: Austrian Academy of Sciences (ÖAW), Vienna Institute of Demography (VID).

Zhang, Z., McDonnell, K. T., Zadok, E., & Mueller, K. (2014). Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map. *IEEE Transactions on Visualisation and Computer Graphics*.

Appendices

Appendix A – Research Proposal

INTRODUCTION

One of the major problems faced by Pakistan since its birth has been the lack of urbanisation and overpopulation in metropolitan cities which has led to impacts that can be dreadful for the country in the coming future if not dealt with accordingly. With the ever-increasing population in Pakistan, the few metropolitan cities such as Lahore, Islamabad and Karachi seem to have increased in size over time to accommodate its inhabitants. The overcrowding in such metropolitan cities has led to increased unemployment, pollution in the cities, traffic and most importantly housing crisis. Considering recent years, the district of Lahore has grown from 263.51 square kilometers in 2000 to 426.8 square kilometers in 2015 almost doubling the previous figure (GHS, 2018). The population in Lahore has also increased from 8.3 million in 2000 to a massive 11.1 million in 2015 (GHS, 2018).

The concept of tackling Urbanisation and population crisis with the use of Big Data Analytics and IoT is something which is still new to the South Asian Countries. Much more fuel is added to this evergreen issue with the increasing number of rural-urban migration in the past few decades. This massive void which is created between these developed metro-cities and the rest of the cities can be addressed by planning new urban areas with the use of modern technology. This paper aims to tackle the Urbanisation and population crisis in Pakistan and its associated issues with the use of Big Data Analytics.

RESEARCH QUESTION:

How to investigate factors contributing towards population and urban crisis in Pakistan using analytical techniques?

RESEARCH AIMS:

The aim of the project is as followed:

Explore the dataset and investigate trends and patterns between population and other various factors that contribute towards it.

Comparing predictive indicators and living standards in different urban cities in Pakistan and other cities in the world.

Conduct deep predictive analysis on how population and urban crisis can be overcome by different analytical solutions.

DELIVERABLE:

A general Big Data report which consists of insights regarding basic demography and factors associated with it. The report outcome will help us determine the factors that contribute towards population and urban crisis in Pakistan and how we can solve this crisis by using analytical techniques.

LITERATURE REVIEW

Considering that Pakistan is the 5th most populated country in the world, it has very few metropolitan cities which can accommodate a decent standard of living to its citizens. In recent years, these metropolitan cities have outgrown in terms of area due to the major housing crisis and development of new housing societies with the lack of planning.

HOUSING CRISIS

Since its independence in 1947, the Governments of Pakistan have been focusing on population control and urban development. However, these developments were mostly development of residential areas without any proper planning. The lack of planning to develop cities is one of the major reasons for the urban crisis that Pakistan is currently going through. The demand and supply of housing in Pakistan has been deteriorating in the past years and continues to do so. The urban demand per year stands at 350,000 units whereas the

supply is only 150,000 (Hasan & Arif, 2018). These statistics lead us to question about the supply and demand of housing in Pakistan considering that the country still has a lot of cities underdeveloped. In past years, the major real estate developments have only played parts in cities like Islamabad, Karachi, and Lahore where residential areas are being developed without any proper planning which has resulted in overpopulation in these cities and thus greater increase per year in size of these cities (Hasan & Arif, 2018). This raises the question as to how the government can tackle this issue using Big Data and IoT to develop other cities.

EMPLOYMENT CRISIS

There are several factors that come into account when discussing the causes of unemployment in Pakistan with population crisis topping the charts. With only a few urban cities such as Lahore, Islamabad and Karachi being the powerhouses to the economy of Pakistan, majority of the citizens must travel to such overly populated urban cities to get employed. Lack of development and resources in other cities also play a major role in migration to such cities for better opportunities. Baluchistan being the largest province of the country has the lowest population as compared to the other provinces speaks much about the Urban Development in Pakistan. However, in recent years the country has started working towards planning and developing new cities in the country to enhance the economy of the country. The CPEC project initiated by the Chinese and Pakistani Governments in Gwadar, has resulted in better opportunities for global trading and tackling unemployment in the country (Nazir, 2021).

In the study conducted by (Cheema & Atta, 2014), it was demonstrated through linear regression model that Unemployment and Population have a positive relationship whereas Inflation and Foreign Direct Investment (FDI) has a negative relationship with Unemployment.

These research studies enable us to question as to how population can be controlled through urban planning using Big Data which could result in achieving lower unemployment rate in Pakistan.

ENVIRONMENTAL ISSUES

In the past few years, the mega cities of Pakistan have at times topped the chart for having the worst Air Quality Index which has resulted in devastating health issues for the general population in the form of smog (a hybrid of Smoke and Fog). Air pollution has developed to be one of the most critical environmental concerns in the 21st Century due to unchecked deforestation, urbanisation, and the rapid growth of industries. The raging urbanisation has led to usage of more vehicles operating mainly on Petrol and Diesel which has led to saturation of polluted particles in the air. Technological advancements in recent years have led us to the innovation of electric cars but Pakistan seems far too away from investing in electric vehicles in its Automobile Industry. Smog has been one of the major factors accounting for increasing cardiovascular diseases in the country. Lack of measures taken by the government has resulted in such alarming situations. Only 1% of the factories in Pakistan report their emissions and there is no penalty or charge as to the CO emitted by vehicles in Pakistan as it is in the western world (Riaz & Hamid, 2018). Lahore and Karachi are the most air polluted cities in Pakistan with an average 190 US AQI and the main pollutant being the PM2.5. The PM2.5 concentration in Lahore is 29.7 times more than the annual air quality guideline set by the World Health Organization (IQAir, 2021).

The challenges to tackle the environmental effects caused by Air Pollution in major cities of Pakistan are numerous. Political ignorance towards the environmental effects in Pakistan counts as one of the biggest challenges. Even though Lahore and Karachi have consistently been topping the charts for the worst air quality index, no proper measurements have been taken by the government to tackle this situation which in future could lead to devastating

health issues in the country. In recent reports, more than 95% of population in South Asia are living in areas where the pollutant concentrations are way over the guidelines set by the World Health Organization. (Anjum, et al., 2021).

The most recent project started by the government of Pakistan has been 'Ten Billion Tree Tsunami Project' to lead the way for the restoration of environment and the ecosystem. The environmental issues and climate changes in Pakistan is due to its large population as 24% of Pakistanis live under the poverty line and are mainly dependent upon natural resources which has led to cases of major deforestation in the country (UNEP, 2021).

Considering this study, a relation can be investigated between the population of the country depending upon natural resources and the environmental factors that comes with it.

HEALTHCARE SYSTEM

The development of healthcare system in Pakistan majorly focuses towards developing Hospitals in big cities such as Lahore, Islamabad, Karachi, Faisalabad, and Rawalpindi. Other cities which are mostly small-scale cities in comparison to the mega cities stated above have inadequate healthcare facilities. In most emergency cases, the patients must be shifted to major hospitals in these metropolitan cities to get better treatment which gets too expensive for the people living under the poverty line. Developed health care system only in the metropolitan cities of Pakistan has led to these cities becoming even more crowded than they already are (Kurji, Premani, & Mithani, 2016).

The lack of healthcare facilities in sub-urban or underdeveloped cities in Pakistan has caused a major rift and loss of many lives in the past couple of years. The healthcare system in Pakistan is administered by the federal government of Pakistan due to which all major health decisions are under their control. Most public health services in Pakistan do not provide adequate services due to which the patients have no other option than to opt for treatment in a private clinic which comes at a very expensive cost. Since 24% of the country's population

lives under the poverty line, it is hard for them to get treated at private hospitals and hence they are left with no other option than to get treated at their local health facilities which are mostly short staffed. In such areas, qualified doctors are hardly available and those who are available mostly recommend the patients to get treated at one of the public hospitals in the metropolitan cities of the country (Kurji, Premani, & Mithani, 2016).

Considering such fragile healthcare system which has only been adequately developed in metro-cities of Pakistan, an investigation can be led to determine why qualified healthcare professionals do not prefer working in cities other than the metro-cities.

EDUCATIONAL CRISIS

One of the most important points that needs to be investigated regarding the population crisis and lack of urban planning revolves around the educational crisis the country faces in most of the cities. Education is the foundation to socio-political, economical development of every country in the world. Education is a basic need for humans to survive and strive in this world. Without the basic level of education, it is very hard to survive in this world as employment and success are heavily dependent upon it.

Since its independence in 1947, Pakistan has been battling with the educational crisis by implementing different kinds of educational reforms to achieve and sustain its educational system. Various reasons play a vital role in the underdevelopment of educational system in Pakistan. Lack of uniformity is one of the major reasons for such a weak education system and presently no such reforms have yet been made to tackle this. In metro-cities we have different kinds of schools such as Public Schools, Private Schools & Deeni Madrassa (Religious Schools). The problem with such system is that public schools have never been up to the mark with only a handful of public universities achieving average rankings around the world. This comes more as a surprise considering that Pakistan has produced Nobel Prize Winners in Dr Abdus Salam, but they have failed to deliver a strong education system to its

inhabitants. The curriculum in schools all over Pakistan are not uniform which has led to private schools teaching different kinds of curriculum such as ones provided by the British Council or The American High School system.

The lack of resources in public schools, colleges and universities is a major reason as to why the students prefer not to study at such institutions where the syllabus is merely updated. The curriculum in such schools does not meet the standards set in current time and follow the same old traditional curriculum. This has massively contributed to students dropping out of schools because of lack of guidance and support by the system. Pakistan is ranked at second when it comes to number of out-of-school children and totals at 22.8 million children aged 5- 16.

Most of the public schools are based in sub-urban areas where the quality of teaching is not up to the mark and the institution is also not equipped with basic equipment. The lack of basic facilities and inadequate training of teaching staff in public institutions can be blamed on the government as the past administrative governments have been allocating less than 2.5% of its annual budget on the education sector (Ahmad, Rehman, Ali, Khan, & Khan, 2014).

The lack of educational institutions outside the metropolitan cities in Pakistan is one of the reasons why students must migrate to such cities to get basic quality education which adds more fuel to the urbanisation and population crisis in Pakistan. This leads to question as to how the population and urbanisation crisis in Pakistan can be tackled by stabilizing the education system of Pakistan with the use of Big Data Analytics.

RESEARCH DESIGN:

Demographic analysis of the population of Pakistan is carried out initially and visualising the data points in a map so that the population distribution of Pakistan can be displayed. Factors like

Exploratory Data Analysis:

In this EDA, we will try to answer the following analysis questions with respect to educational experience:

Exploratory data analysis is carried out on the Pakistan census data before merging the supplementary datasets like educational, real estate, employment etc.

How is the population in urban centres different from rural areas.?

This demographic data will be preprocessed, analysed, mapped to display the underlying patterns so that the policies can be drafted. Towards the end the data will be used to construct a model for predicting child poverty in other cities which will be tested on the same census data variables.

DATA COLLECTION:

The primary demographic data is collected and gathered from the Pakistan census website. The data is ranging from 1998 to 2017. The sample size has been set at around 17600 households, divided into 1252 sample villages/enumeration blocks for provinces in most of the surveys (Methodology | Pakistan Bureau of Statistics, 2020). On a monthly and yearly basis, it collects data from numerous sources, such as official records of federal ministries, departments, and organisations, and provincial governments for the dataset. The secondary data sources are from various departments from employment, education, and housing. The obtained data is in .pdf format and is converted into .csv format for further analysis.

TOOLS USED:

OBJECT ORIENTED LANGUAGE: Python 3

PLATFORM: Jupyter Notebook

PACKAGES:

NumPy- for numerical and mathematical computation

Pandas- for data manipulation and processing, matplotlib- for data visualisation and plotting geographic data.

PROPOSED METHODOLOGY:

The whole project is based on finding the factors which contribute to the rapid urbanisation and devising policy around it to control the crisis behind them. After data collection, the data sets are merged into a single dataset so that the evaluation can be done on a mutual scale (ThaliaTEST SilverTEST, 2018). The merged dataset would have a few constraints such as geographical and historical constraints. For instance, Khyber Pakhtunkhawa (KPK) was named NWFP before and would need to be corrected before we start our analysis. Mostly, exploratory data analysis is performed on the dataset to discover keyful insights. As the collected data is in categorical and numerical form, binary classifiers for different problem statements can be easily identified (Alsharkawi et al., 2021). The dataset is visualised by the latitude and longitude features and the

population is visualised using a bubble plot where the size of the bubble denotes the overpopulation. As we collect the data from 1998 some of the data is missing and could be deemed invalid in our case. So, this missing value is dealt by deleting

some invalid rows and filling up values through multiple imputation accordingly (Baraldi & Enders, 2010). The type of missing value is identified and filled accordingly. If the data is MNAR (Missing Not at Random), the rows should be deleted because it is inappropriate for imputation (Aude et al., 2020). This contributes towards just one area of the data cleaning process that would be applied on our collected data to make sure our analysis is efficient and accurate. Data normalisation is the next step carried after the data cleaning process. Standardisation of the data is important because some variables would have a huge value such as 'Population' whereas other variables would be having very less numerical value (Collins et al., 2001). Hence to analyse and compare these variables on a mutual scale, we would have to standardise our data. Population, being a huge numeric value, would be viewed on a log scale. Pearson correlation coefficient is used to find the relationship between all the variables (Taylor, 1990). Descriptive statistics is carried out on the features to generate overview about the dataset. Multiple linear regression is used for predictive analysis of factors which have contributed towards the urbanisation and population crisis in Pakistan.

MULTIPLE LINEAR REGRESSION:

Independent variables and dependent variables are identified in our case, population is considered as the dependent variable and various factors like employment rate, literacy rate, inflation rate is considered as the independent variables. The independent variables are identified as per the descriptive statistics of the dataset. The impact of the independent variables on the dependent variable is identified. Confusion matrix is used for evaluating the multiple linear regression.

ETHICS:

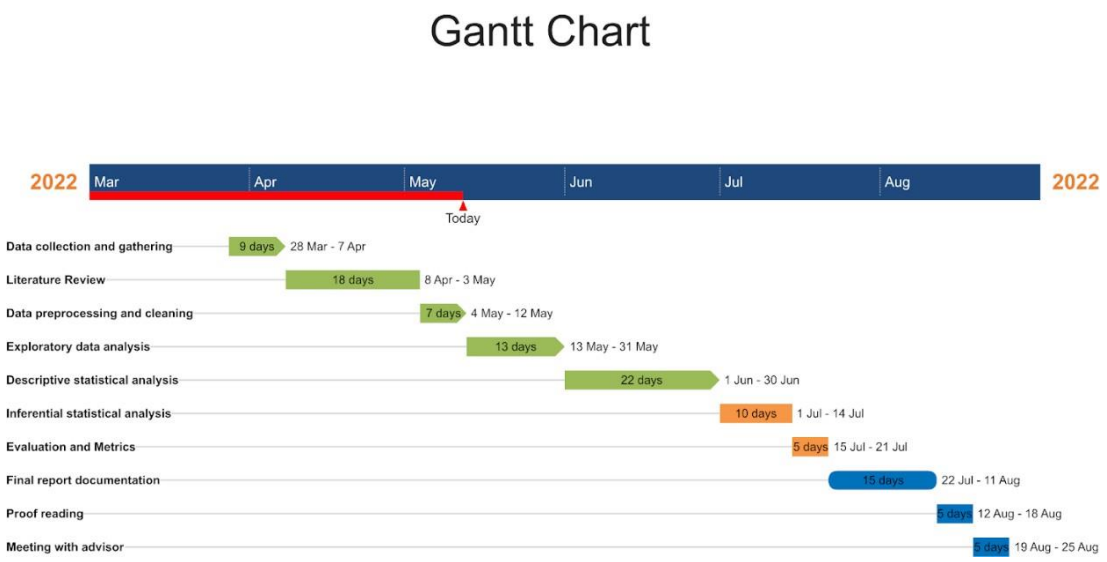
The dataset is a collection of information that the Pakistan Bureau of Statistics has provided to demonstrate the platform's transparency. Pakistan Bureau of Statistics has provided data sets regarding population and census to the public. The dataset is considered ethical because the platform has made it available and using it for academic purposes is an ethical practice. As a result, no ethical approval is required for this secondary source of data. The data provided in the dataset by the Pakistan Bureau of Statistics is collected by the bureau and is being used as a secondary dataset in this project.

RISKS:

Census data is not uniform throughout the years as the format and the accuracy is different in different time periods which might affect the efficiency of the machine learning model. Inaccurate or incomplete data can be poised as a risk considering the scope of this project and would need to be carefully cleaned to apply the desired analytical techniques on them. While

merging the datasets, we must make sure that every variable is named correctly without any errors so that proper relationships between variables can be easily identified and analysed.

GANTT CHART:



REFERENCES

Ahmad, I., Rehman, K. u., Ali, A., Khan, I., & Khan, F. A. (2014). Critical Analysis of the Problems of Education in Pakistan:. *International Journal of Evaluation and Research in Education (IJERE)*, 79-84.
Anjum, M. S., Ali, S. M., Imad-ud-Din, M., Subhani, M. A., Anwar, M. N., Niazmi, A.-S., . .

. Khokhar, M. F. (2021). An Emerged Challenge of Air Pollution and Ever-Increasing Particulate Matter in Pakistan; A Critical Review. *Journal Of Hazardous Materials*.
Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., & Alyaman, M. (2021). Poverty Classification Using Machine Learning: The Case of Jordan. Sustainability, 13(3), 1412. https://doi.org/10.3390/su13031412

Aude, S., Claire, B., & Julie, J. (2020). Estimation and Imputation in Probabilistic Principal Component Analysis with Missing Not At Random Data. Advances in Neural Information Processing Systems, 33.
https://proceedings.neurips.cc/paper/2020/hash/4ecb679fd35dcfd0f0894c399590be1a-Abstract.html

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of School Psychology, 48(1), 5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330–351. https://doi.org/10.1037/1082-989x.6.4.330

GHS. (2018, June). *Built up area of Lahore in Pakistan from 1975 to 2015*. Retrieved from www.statista.com: https://www.statista.com/statistics/911670/pakistan-built-up-area- lahore/
GHS. (2018, June). *Number of inhabitants in Lahore in Pakistan from 1975 to 2015*.

Retrieved from www.statista.com: https://www.statista.com/statistics/911007/pakistan-population-in-lahore/
Hasan, A., & Arif, H. (2018, October). *Pakistan: the causes and repercussions of the housing crisis*. Retrieved from NYU Faculty Digital Archive: http://hdl.handle.net/2451/44207
Kurji, Z., Premani, Z. S., & Mithani, Y. (2016). Analysis of the health care system of Pakistan: lessons learnt and way forward. *eCommons AKU*.
Riaz, R., & Hamid, K. (2018). *Existing Smog in Lahore, Pakistan: An Alarming Public Health Concern.* Cureus.
UNEP. (2021, June 02). *Pakistan's Ten Billion Tree Tsunami*. Retrieved from United Nations Environment Programme: https://www.unep.org/news-and-stories/story/pakistans-ten- billion-tree-tsunami

Cheema, A. R., & Atta, A. (2014). Economic determinants of unemployment in Pakistan: Co-integration analysis. *International journal of business and social science*, *5*(3).
Nazir, H. (2021). Impact Assessment of Gwadar Port on China-Pakistan Economic Corridor: A Case Study. *Journal of Art, Architecture and Built Environment*, *4*(1), 69-95. https://doi.org/10.32350/jaabe.41.04

IQAir (2021). Air Quality in Lahore. Retrieved from https://www.iqair.com/pakistan/punjab/lahore

Methodology | Pakistan Bureau of Statistics. (2020). Pbs.gov.pk. https://www.pbs.gov.pk/content/methodology-1

Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. Journal of Diagnostic Medical Sonography, 6(1), 35–39. https://doi.org/10.1177/875647939000600106

ThaliaTEST SilverTEST. (2018). How to analyse quantitative data for evaluation — NCVO Knowhow. Ncvo.org.uk. https://knowhow.ncvo.org.uk/how-to/how-to-analyse-quantitative-data-for-evaluation

Sheffield Hallam University

UREC 1 RESEARCH ETHICS REVIEW FOR STUDENT RESEARCH WITH NO HUMAN PARTICIPANTS OR DIRECT COLLECTION OF HUMAN TISSUES, OR BODILY FLUIDS.

All University research is required to undergo ethical scrutiny to comply with UK law. The University Research Ethics Policy (https://www.shu.ac.uk/research/excellence/ethics-and-integrity/policies) should be consulted before completing the form. The initial questions are there to check that completion of the UREC1 is appropriate for this study. The supervisor will approve the study, but it may also be reviewed by the College Teaching Program Research Ethics Committee (CTPREC) as part of the quality assurance process (additional guidance can be obtained from your College Research Ethics Chair[1])

The final responsibility for ensuring that ethical research practices are followed rests with the supervisor for student research.

Note that students and staff are responsible for making suitable arrangements to ensure compliance with the General Data Protection Regulations (GDPR), for keeping data secure and if relevant, for keeping the identity of participants anonymous. They are also responsible for following SHU guidelines about data encryption and research data management. Guidance can be found on the SHU Ethics Website https://www.shu.ac.uk/research/excellence/ethics-and-integrity

Please note that it is mandatory for all students to only store data on their allotted networked drive space and not on individual hard drives or memory sticks etc.

The present form also enables the University and College to keep a record confirming that research conducted has been subjected to ethical scrutiny. Students should retain a copy for inclusion in their research projects, and a copy should be uploaded to the relevant module Blackboard site.

The form must be completed by the student and approved by supervisor and/or module leader (as applicable). In all cases, it should be counter-signed by the supervisor and/or module leader and kept as a record showing that ethical scrutiny has occurred. Students should retain a copy for inclusion in the appendices of their research projects, and a copy should be uploaded to the module Blackboard site for checking.

Please note that it may be necessary to conduct a health and safety risk assessment for the proposed research. Further information can be obtained from the University's Health and Safety Website

1. General Details

Details	
Name of student	Tallal Ahmed Bhatti
SHU email address	C1042155@my.shu.ac.uk
Department/College	Computing Department
Name of supervisor	Dani Papamaximou
Supervisor's email address	D.Papamaximou@shu.ac.uk
Title of proposed research	Tackling Urbanization and population issues in Pakistan using analytics
Proposed start date	1/06/2022
Proposed end date	08/09/2022
Brief outline of research to include, rationale (reasons) for undertaking the research & aims, and methods (250-500 words).	The research revolves around investigating the population and urbanization crisis in Pakistan. I believe that the overpopulation in the metropolitan cities of Pakistan lays foundation for many of the country's problems. A country with such a high population should at least have a decent population density all over the land. The research is conducted to finding answers to the issues by exploring relations between variables and what efforts can be done to solve this crisis.

I confirm that this study does not involve collecting/using data or samples from human participants

Please tick **YES**

2. Research in external organizations

Question	**Yes/No**
1. Will the research involve working with/within an organization (e.g., school, business, charity, museum, government department, international agency, etc.)?	YES
2. If you answered YES to question 1, do you have granted access to conduct the research? *If YES, students please show evidence to your supervisor. PI should retain safely.*	YES

3. If you answered NO to question 2, is it because: A. you have not yet asked B. you have asked and not yet received an answer C. you have asked and been refused access. *Note: You will only be able to start the research when you have been granted access.*	

3. **Research with Products and Artefacts**

Question	**Yes/No**
1. Will the research involve working with copyrighted documents, films, broadcasts, photographs, artworks, designs, products, programs, databases, networks, processes, existing datasets, or secure data?	YES
2. If you answered YES to question 1, are the materials you intend to use in the public domain? *Notes: 'In the public domain' does not mean the same thing as 'publicly accessible'.* • *Information which is 'in the public domain' is no longer protected by copyright (i.e., copyright has either expired or been waived) and can be used without permission.* • *Information which is 'publicly accessible' (e.g., TV broadcasts, websites, artworks, newspapers) is available for anyone to consult/view. It is still protected by copyright even if there is no copyright notice. In UK law, copyright protection is automatic and does not require a copyright statement, although it is always good practice to provide one. It is necessary to check the terms and conditions of use to find out exactly how the material may be reused etc.* *If you answered YES to question 1, be aware that you may need to consider other ethics codes. For example, when conducting Internet research, consult the code of the Association of Internet Researchers; for educational research, consult the Code of Ethics of the British Educational Research Association.*	YES
3. If you answered NO to question 2, do you have explicit permission to use these materials as data? *If YES, please show evidence to your supervisor.*	
4. If you answered NO to question 3, is it because: A. you have not yet asked permission B. you have asked and not yet received and answer C. you have asked and been refused access. *Note You will only be able to start the research when you have been granted permission to use the specified material.*	**A/B/C**

4. **Does this research project require a health and safety risk assessment for the procedures to be used?** Discuss this with your supervisor and consult the Risk Assessment Toolkit for teaching research.

Yes
No

(If **YES** the completed Health and Safety Risk Assessment form should be attached). You can find a Blank/Sample Risk Assessment Form at the Checklist, Generic and TORS Risk Assessments on the Risk Assessment Toolkit

Adherence to SHU policy and procedures

Ethics sign-off	
Personal statement	
I can confirm that: • I have read the Sheffield Hallam University Research Ethics Policy and Procedures • I agree to abide by its principles.	
Student	
Name: TALLAL AHMED BHATTI	Date: 24/08/2022
Signature: TALLAL AHMED BHATTI	
Supervisor or another person giving ethical sign-off	
I can confirm that completion of this form has confirmed that this research does not involve human participants. The research will not commence until any approvals required under Sections 2 & 3 have been received and any health and safety measures are in place.	
Name:Dani Papamaximou	Date: 14/09/2022
Signature:DANI PAPAMAXIMOU	
Additional Signature if required:	
Name:	Date:
Signature:	

Please ensure that you have attached all relevant documents. Your supervisor must approve them before you start data collection:

Relevant Documents	**Yes**	**No**	**N/A**
Research proposal if prepared previously			
Any associated materials (e.g., posters, letters, etc.)			
Health and Safety Risk Assessment Form			

Appendix C – Publication Form

Sheffield Hallam University | College of Business, Technology and Engineering

Research Skills and Dissertation Module (55-706556).

PUBLICATION PROCEDURE FORM

In this module, while you create your own research question or topic area, your supervisor makes a significant intellectual contribution to this work as the research progresses. Your supervisor will make the decision on whether your work merits publication based on the quality of the work you have produced. Your supervisor will co-author the paper for publication with you and your supervisor will both be listed as authors. You are required to sign the declaration below to confirm that you understand and will follow this procedure.

Declaration:

I, Tallal Ahmed Bhatti, confirm that I understand will comply with the Publication Procedure outlined in the Module Handbook and the Blackboard Site.		
Student:	Signature Tallal Ahmed Bhatti	Date 15.09.22
Supervisor:	Signature Dani Papamaximou	Date 15/09/2022

Appendix D – Data documents and Information

Datasets Used:

Pakistan Bureau of Statistics Census 2017 (https://www.pbs.gov.pk/content/final-results-census-2017-0):

1. Population Data – Table 01 of Census report
2. Employment Data – Table 16 of Census report
3. Education Data – Table 14 of Census Report
4. Housing Data – Table 30 of Census Report

Al- Hasan Systems Initiative for Open Data (https://data.humdata.org/dataset/pakistan-health-facilities)

5. Health Data – All health data with shapefiles from the year 2012

GitHub Link containing all material relevant to research in the branch (Dissertation-Submission):
https://github.com/lordtallal/Dissertation/tree/Dissertation-Submission