

**TASK**

**Exploratory Data Analysis on the Commuting Time Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Millions of people commute to business centres daily, studies have shown the various consequences this has on productivity and morale. On this task we have commuting time dataset obtained and captured from one of South Africa’s Provinces [Western Cape].

The Dataset captured the various information per household with a unique household Identifier, different income and age groups and ethnicity.

On this dataset we are going to look at the time taken in commuting, correlation between variables. With the given dataset we can come up with the correlation between Income and age, race and probably the level of education. Furthermore, we will be able to see how much time it takes in commuting. To understand the dataset, there is need to iterate over the variables individually and get an understanding.

**DATA CLEANING**

Data Cleaning is a process of fix bad data in any given set, this includes missing /empty cell, data in wrong formats, duplicates. All this can potentially give us biased data when we run the analysis.

On the Commuting dataset I used various techniques to clean the data. Firstly, the main goal is to identify and create relationship between variables that helped in coming up with ideas. I had to remove some columns using the drop function, these columns add no value to what I aim to analyse from the data set. Secondly, had remove the 2 duplicate entries

Checking the data types is one of the most important steps in cleaning data as it helps in further decision making. It is important to make sure that data is consistent. Much of the data was classified as objects. Conversion of data types for easier analysis.

The age category has too much data to go through and had to be grouped in age groups of 10 years.

MISSING DATA

Analysing missing data would give biased result, therefore dealing missing data is variant on the magnitude.

A picture containing text, monitor, screen, screenshot

Description automatically generated

As shown on the plot above only one column showed missing data and used the drop function to delete from the dataset.

The Dataset comprised of invalid/null data entries [Not Applicable and Unspecified] which refer to missing data. To rectify that we used the replace function and imputation, such that we can derive some meaning and convert the data type for analysis.

DATA STORIES AND VISUALIZATIONS

To understand the data, had to summarize the different characteristics and finding representative/critical points for analysis. The process was done on individual variables to get insight as shown below:

We have more females than males, sorting the data according to race was able to figure that the population leaned more towards the coloured community which is understandable considering the province of reference.

Shape, rectangle

Description automatically generated

Icon

Description automatically generatedFurthermore, urban locations have more people that commute to work compared to those from metro and rural as shown below:

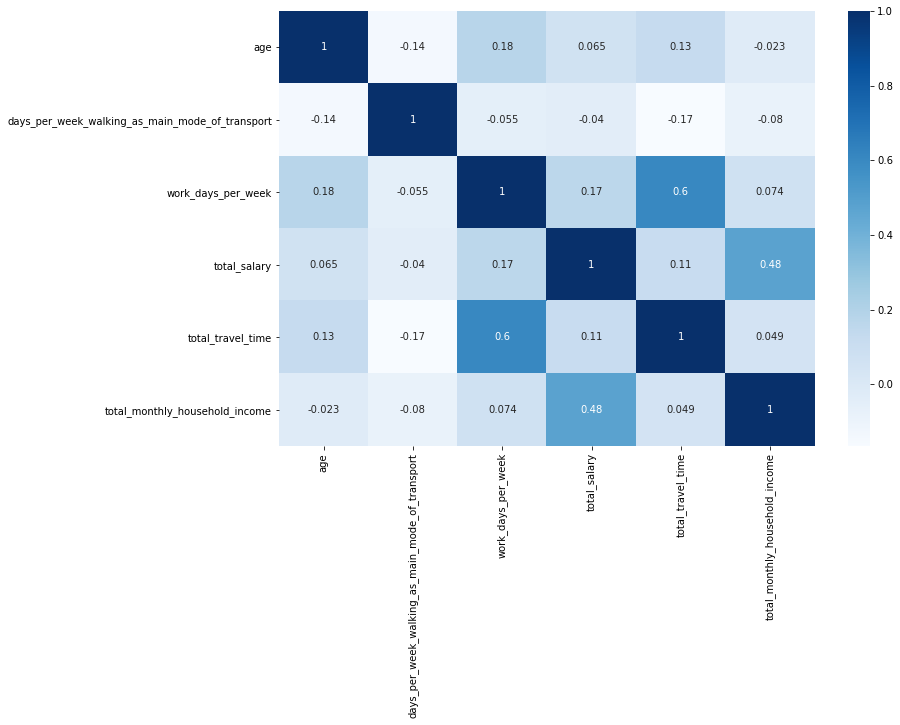
Majority of the population is not employed, with informal sector employing a

smaller percentage

Chart, bar chart

Description automatically generated

With the given dataset aim was to figure out the correlation between the variables in so doing, plot below shows the correlation that exists on the variables in question:

  
  
To add on , had to plot the individual variable to further show the lack of correlation on the give data;There is no correlation between total salary earned/ total monthly income per unique household and the age.

A picture containing shape

Description automatically generated

Chart, scatter chart

Description automatically generated

Does race, age and location influence the commuting times?

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Conclusively, the data set chosen has a few anomalies and some inferences didn’t come out clearly

**THIS REPORT WAS WRITTEN BY: BLISS J MUSISINYANI**

