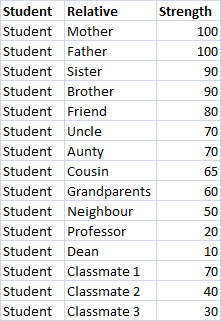
**PART I : Chord Diagram of student and 15 other persons**

A force-directed graph, chord diagram helps to understand the relationship between the student and 15 people.

The graph represents:

1. The relationship between the student and the person
2. The strength of the relationship

We have created a relationship table between the student and a person and associated with a value i.e. strength. A random data is created. Data includes the student and 15 other persons and the strength of the relation. To understand the relationship through visualization I’m using the tool Power Bi. The data is imported to power Bi and chord diagram is generated.



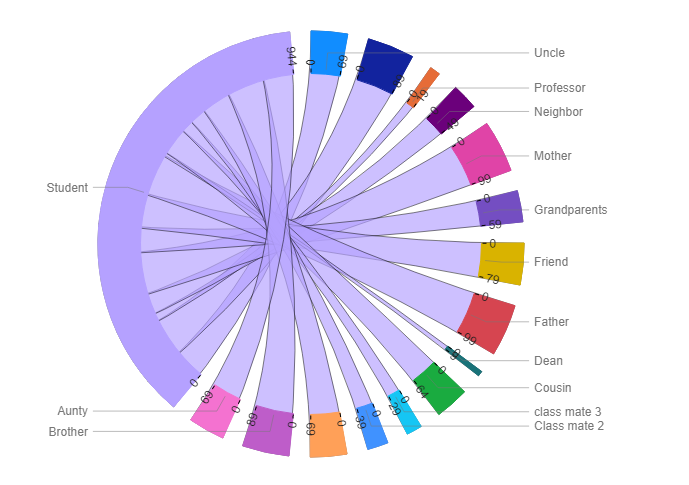
****

Fig 1: Relationship diagram of student and persons

Chord diagram represents the interrelationship visualization in the form of matrix. The representation is in the form of circle and length of the arc is the strength of the relationship. The data points are connected in the form of arc and link represents they share a common entity in a group. Greater the strength of the arc stronger the relationship. Here we have used different color to represent the relative and to distinguish other relation.

The above chord diagram is generated from the student relationship data which was created to understand the student and association with 15 other relatives. From the diagram, The length of the arc for Father and Mother is more compared to other relatives as the strength is 100, and the length of arc is least with the dean as the strength is 10.

**Part II: EDA on Crime Data**

Crime Dataset includes the details of crime details like case number, ICUR, Domestic, Arrest, FBI code, what is the primary type, location of crime, year in which crime is committed and many other details and there are around 6 million crime cases details.

In the given dataset case number was a dimension and for the analysis of the crime data the case number is converted into measure.

Exploratory analysis on Crime Data:

* **Details of FBI Codes.**

In the given crime data set, each case number is associated with the FBI code. There are around 26 distinct FBI codes.

CNT(Case Number)

Sum: 5,862,792

Average: 225,492.00

Minimum: 22

Maximum: 1,211,080

Median: 61,168.50

CNTD(Case Number)

Sum: 5,862,508

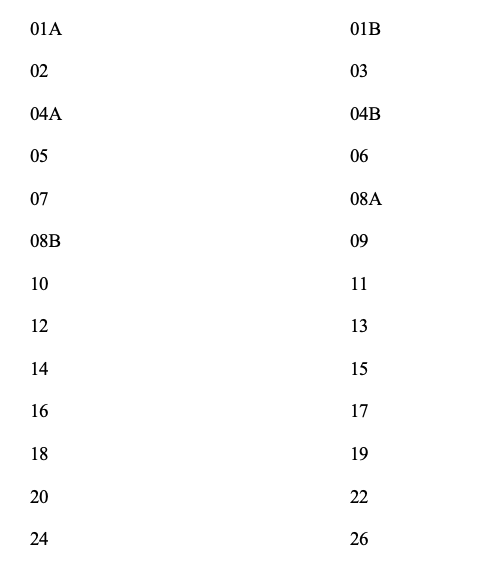
Average: 225,481.08

Minimum: 22

Maximum: 1,211,078

Median: 61,168.00

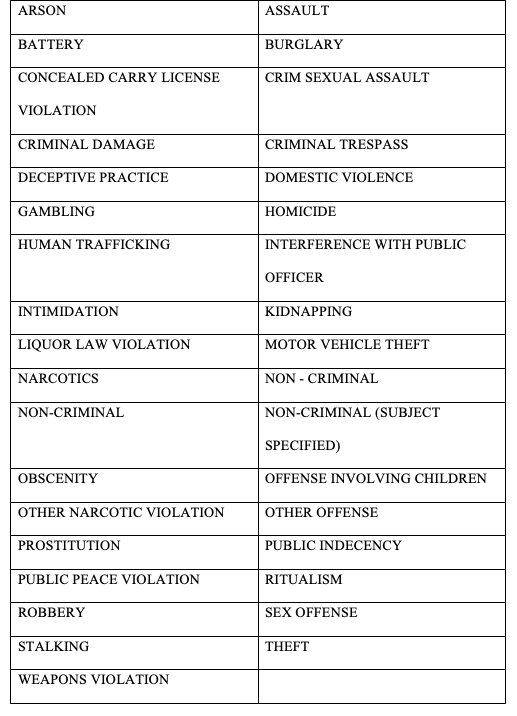
FBI Codes:



|  |  |
| --- | --- |
|  |  |

* **Details of primary types of crime.**

In the given crime data set, each case number is associated with the primary type of crime. There are around 34 distinct primary types of crime.



* **Total number of criminal cases v/s FBI code**

From the given crime data, almost all the 6 million crime records are associated with the either of the 26 distinct FBI codes.



Fig. 2 - FBI code v/s count of case number

To analyze the number of crime cases related to each FBI code, a bar graph is used as we compare the total cases number of cases every easily. In the Fig. 2, we can see the total number of criminal cases recorded for each FBI code from crime\_data set.

From the above bar chart, the total number of records v/s the FBI codes is in the decreasing order. We can see that the number cases related to FBI code “06” is highest i.e. 1,211,080 and the total number of cases related to FBI code “01B” is lowest. i.e. just 22 cases. In the given data set, there are 26 distinct FBI code and around 6 million cases and each case is associated with a FBI code. We may not know which type of crime is committed by looking at the FBI codes.

* **Total number of criminal cases v/s primary type.**

From the given crime data, almost all the 6 million crime records out of which there 5,862,508 distinct criminal records and are associated with the either of the 34 distinct primary type.



Fig. 3 - primary type v/s count of case number

To understand the number of criminal cases associated with each of the 34 primary types, a bar graph is used as we compare the total cases number of cases every easily. In the Fig. 3, the criminal cases are related to either of the primary cases. The bar chart is in the ascending order of total number of criminal cases. The highest number of cases recorded for the type “Theft” i.e. 1,211,078 and the lowest number of cases recorded for the type “Domestic Violence” i.e. just 1 record. From the given data

* **Comparison of crime records with both primary type and FBI code.**

Each criminal record is associated with FBI code and the primary type. We want to compare which of type the crime is most committed and associated FBI code.



Fig. 4 ﻿Criminal cases v/s FBI code and Primary Type.

The above Tree map displays the total number of cases related to primary type and the associated FBI code of the case in the descdending order. From the map, we interpret that the type of crime “Theft” has the highest number of cases i.e. 1,211,080 associated with the FBI code “06” . Next to type “Theft”, “Battery” is having more number of cases i.e. 916,477 associated with the FBI code “0B” and then “Criminal Damage” associated with the FBI code “14”.

In the **Tree map** the information is represented in the form of reactangle. Larger the area of the rectangle highest is the number of the cases. Based on the dimensions of the rectangle the color and shade of the rectangle varies. The dark green color indicates the highest number of cases where as dark red represents the least number of cases

* **Time frame of the crimes committed.**

To understand the total number of cases over a period of time we used line chart, as it gives us the clear picture of trend over the time period.



Fig. 5 - Time frame of the crimes committed

The line chart helps to represent the data associated with the time. So to understand the count of criminal cases during the time period 2001 to 2015 is represented using the line chart.

The above line chart shows the data that the 5,862,795 cases information is gathered during the time frame 2001 to 2015. From the graph, We can see the highest number i.e. 486,703 crimes committed during the year 2002. The total number of the criminal cases committed are decreasing over the period from 2002 to 2015. The least number of cases i.e. 151,632 are in the year 2015. Almost 68.9% of reduction in the total number of crimes in comparison with total count of cases in the year 2002.



Fig. 6 – Trend of change in the total crimes committed.

The above graph helps to understand in detail how the change in the total count of cases year by year. We see there is drop in the cases in compare to the year 2002. From the year 2003 to 2008 there is increase and decrease of cases compare to previous year. From the year 2008 to 2012, continuously the criminal cases increased. Suddenly there is huge drop in the total cases in the year 2015.

* **In Detail trend of the total crimes committed in association with FBI codes and Primary type**.

To understand the trends of the crime committed in associated with FBI codes and primary type i use area charts to get the clear picture of which year more number of crimes committed.



In the graph, the primary type is distinguished using the different colour. We see that the in the year 2006 most the criminal cases i.e. around 3,904 are related to the type “Arson” and only single case is recorded for the type “Battery” in the year 2001. Initially the trend is low, but during the year 2006, the highest number of cases are updated in the criminal record. From the year 2007 the cases updated in each year is slowing decreasing but we can a spike in the year 2014 and then decreased in the year 2015. The Primary types of crimes that were committed initially are “Battery” , “Narcotics”, “Liquor Violation” until the year 2006. Later on many other different types of crimes evolved over the period.

**Part III: Data visualization of Sales Record**

The sales information is collected over the period from 2014 to 2017 across all the multiple regions in the US. The sales information includes profit, discounts, shipment date and shipment mode and many other attributes information is collected.

* **Understanding of sales and profits gained w.r.t each different category and each different region**

From the given sales information, to understand the sales and profits made w.r.t category and region, we use the scatter plot. To visualize the relationship between numeric variables i.e. between independent and the dependent variables.



Fig.7 sum of profit vs sum of sales w.r.t Region and Category

The above scatter plot is to understand the relationship between the sum of sales and the sum of the profits. Here in the scatter plot sales is the dependent numeric variable and profits in the independent numeric variable. To understand the how the total number of sales in different regions, I have distinguished using the different colour and to understand the sales information with category I have used different shapes.

From the above scatter plot, the total profits and total sales are less in the South Region in comparison with other regions. Overall as the total profits increase the number of sales are also increase. In the region West, the profits are less even though the number of the sales are more. In all the regions, the Technology category is having highest number of sales and the profits. In all the regions, the profits are less in the Furniture category but there is no great difference between in the total of sales they are significant w.r.t Technology and Office supplies category.

﻿Count: 12

SUM(Profit)

Sum: 286,397

Average: 23,866

Minimum: -2,871

Maximum: 52,610

Median: 19,989

SUM(Sales)

Sum: 2,297,201

Average: 191,433

Minimum: 117,299

Maximum: 264,974

Median: 187,966

The business insight that can be inferred by looking at the above scatter plot is in all the regions Technology category is having highest number of sales and profit. And also the office supplies category are contributing major sales in comparison with the Furniture category.

* **Understanding of sales and profits gained w.r.t each segment, each different category and each different region.**

To understand the sales information, in each segment, in each category and in each region we use the scatter plot. As scatter plot visualizes the relationship between the two variables. How one variable is corelated with another variable. Here understand how the sales are corelated with the profit scatter plot can gives the clear picture.



Fig. 8 - sum of profit vs sum of sales w.r.t Segment, Region and Category

From the above Fig. 8, the scatter plot shows the correlation between the sales and the profit.

Here, to get the clear understanding the Region are distinguished using the colour, the segments are distinguished using the shape and the category is distinguished using size.

In the Central region, all the segments mainly in the furniture category the profits are negative. We see that the correlation between the sales and the profits are directly proportional. The highest number of sales and the profits are made in the consumer segment in the Technology category, where lowest number of sales are made in the South region, in the Home office segment in almost all the categories. In the Corporate segment even though we see the significant number of sales but the profits are less.

﻿Count: 12

SUM(Profit)

Sum: 286,397

Average: 23,866

Minimum: -2,871

Maximum: 52,610

Median: 19,989

SUM(Sales)

Sum: 2,297,201

Average: 191,433

Minimum: 117,299

Maximum: 264,974

Median: 187,966

The business insight from the above scatterplot is, as we have seen negative profits in the furniture category in mostly all the segments. The business need to look for options like discounts to increase the sales and the profits. In the Technology category, consumer segment is making maximum profit so continue to produce the better quality products and increase the sales.

**Part IV: EDA on pima diabetes data**

To understand the pima diabetes data we perform EDA on the data. Pima diabetes data is collected with based on the 9 different attributes.

We understand the data by visualizing histogram, Density plot and Boxplot.

* **Plot histogram of each numeric variable / attribute in the data set**

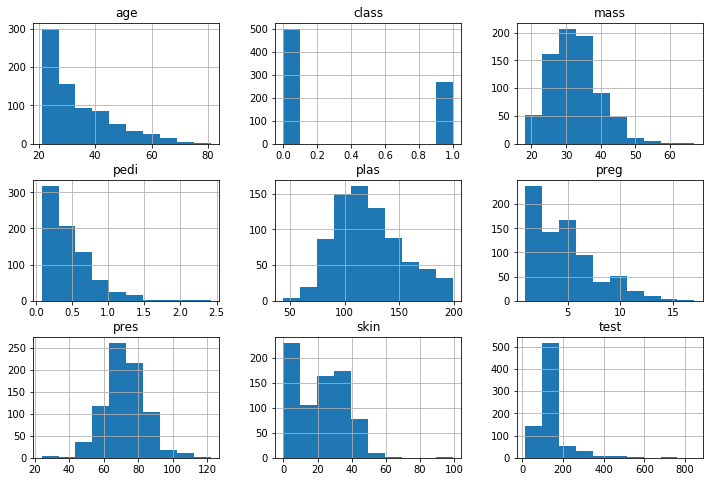


Fig. 10 Plot histogram of each numeric variable / attribute in the data set

* **Density plots of each numeric variable / attribute in the data set**

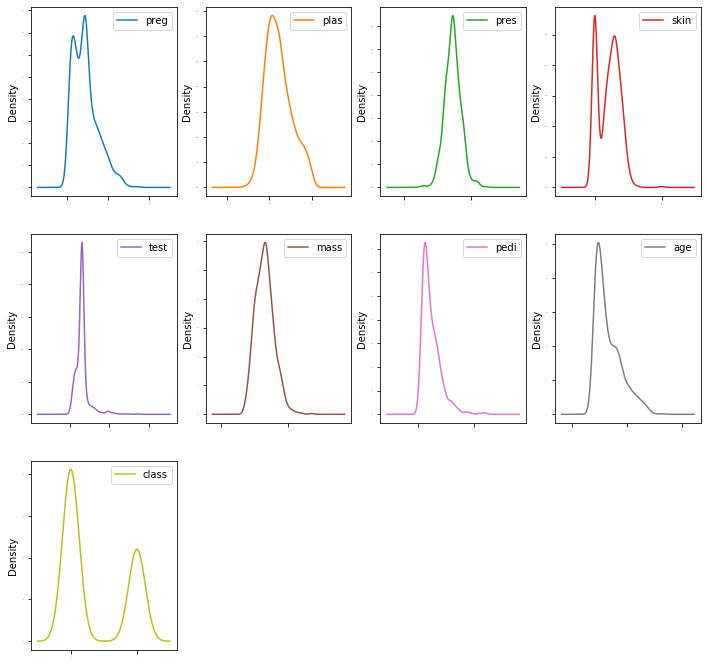
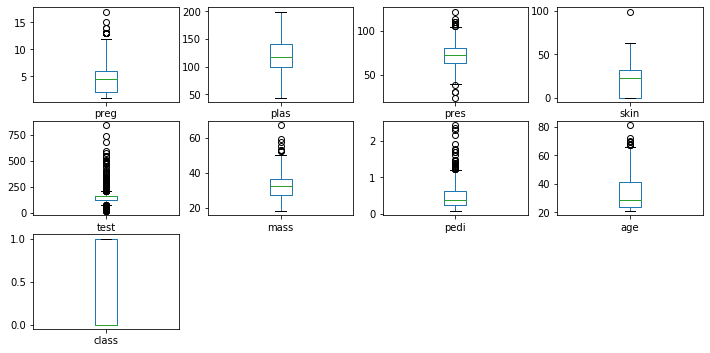


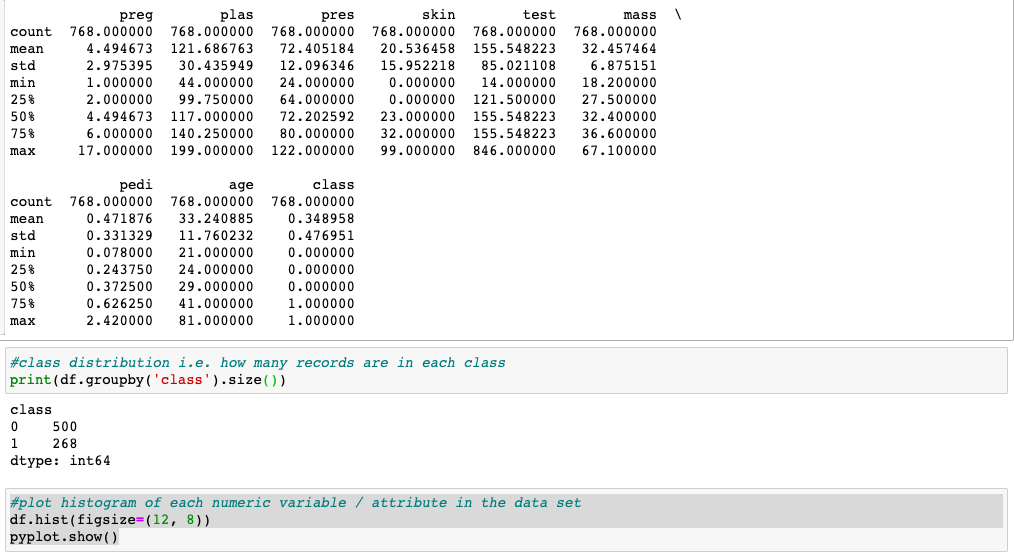
Fig. 11 - Density plots of each numeric variable / attribute in the data set.

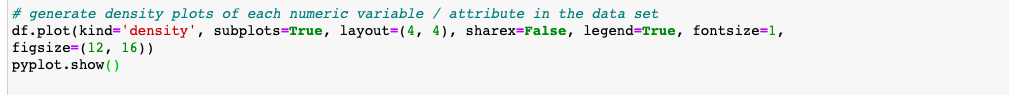
* **Box plots of each numeric variable / attribute in the data set**

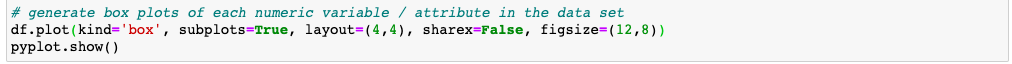


* **Python code and results to understand the Pima diabetes data.**

****

****

****

****

﻿