**SP24: DATA-245 Sec 11 – Machine Learning Tech**

**Homework - - 1**

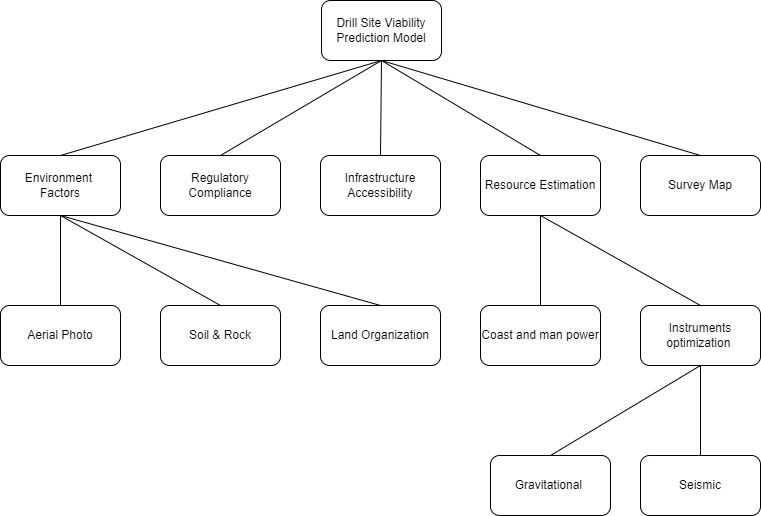
**Name :- Prayag Nikul Purani**

**SJSU Id :- 017416737**

**Question 1:-**

Chapter 2, exercise 7

1. The prediction subject that can be used for this situation can be potential drilling site.
2. Finding a suitable location for drilling depends on a number of elements, including the sort of drilling you're thinking about (oil, gas, or water), as well as the local geology. Examine geological data and maps to comprehend the makeup and structure of the region's subsoil. Determine whether prospective formations or reservoirs could hold the needed resources. To learn more about underlying structures and locate possible reservoirs, conduct seismic studies. To find subsurface features, use gravity and magnetic surveying techniques. Examine how drilling would affect the environment in the selected location, taking into account surrounding waterways, ecosystems, and populated areas. Make sure that all local, state, and federal laws pertaining to drilling operations are followed. Obtain the required authorization and permissions from the appropriate authorities. Examine the site's accessibility with regard to pipelines, roads, and other infrastructure requirements. Based on the information at hand, calculate the possible yield of the resources you are drilling for (such as water, natural gas, or oil reserves). Do a cost-benefit analysis, weighing the possible returns against the expenses of drilling, exploration, and operations. Consider possible dangers including earthquake activity, geological instability, and other hazards. Create strategies for emergency response and safety precautions. Interact with the community to learn about any issues they may be facing and resolve them. Provide an open line of contact with stakeholders to ensure they are aware. Examine if it is feasible to construct the infrastructure needed for resource extraction and transportation.



1. It is unlikely that any particular legal issues would arise from the data used for this scenario.

**Question 2:-**

Chapter 3, exercise 5

1. Min = 27, max = 89, low = 0, high = 1



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| score | 0.24 | 0.32 | 0.52 | 0.00 | 0.92 | 0.35 | 0.73 | 0.26 | 0.74 | 0.52 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| score | 0.50 | 0.89 | 0.37 | 0.84 | 1.00 | 0.77 | 0.69 | 0.52 | 0.65 | 0.13 |

1. Min = 27, max = 89, low = -1, high = 1



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| score | -0.52 | -0.35 | 0.0. | -1.00 | -0.84 | -0.29 | -0.45 | -0.48 | 0.48 | 0.03 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| score | 0.00 | 0.77 | -0.26 | 0.68 | 1.00 | 0.55 | 0.39 | 0.03 | 0.29 | -0.74 |

1. Mean = 60.95 , std = 17.25



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| score | -1.10 | -0.81 | -0.11 | -1.97 | 1.34 | -0.69 | 0.64 | -1.04 | 0.70 | -0.11 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| score | -0.17 | 1.22 | -0.63 | 1.05 | 1.63 | 0.81 | 0.52 | -0.11 | 0.35 | -1.50 |

**Question 3:-**

Chapter 3, exercise 6

1. Equal-width binning using 5 bins

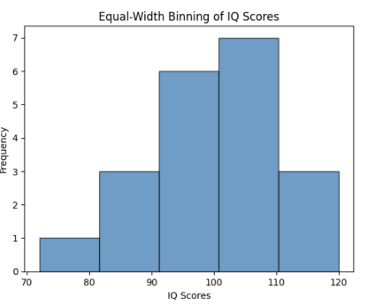
Range = (high - low)/ no. of bins

Range = (120-72)/5

= 9.6

|  |  |  |
| --- | --- | --- |
| Bin no. | Low | High |
| 1 | 72.0 | 81.6 |
| 2 | 81.6 | 91.2 |
| 3 | 91.2 | 100.8 |
| 4 | 100.8 | 110.8 |
| 5 | 110.4 | 120.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | IQ | Bin no |  | ID | IQ | Bin no |
| 1 | 92 | --3 |  | 11 | 105 | --4 |
| 2 | 107 | --4 |  | 12 | 88 | --2 |
| 3 | 83 | --2 |  | 13 | 106 | --4 |
| 4 | 101 | --4 |  | 14 | 90 | --2 |
| 5 | 107 | --4 |  | 15 | 97 | --3 |
| 6 | 92 | --3 |  | 16 | 118 | --5 |
| 7 | 99 | --3 |  | 17 | 120 | --5 |
| 8 | 119 | --5 |  | 18 | 72 | --1 |
| 9 | 93 | --3 |  | 19 | 100 | --3 |
| 10 | 106 | --4 |  | 20 | 104 | --4 |

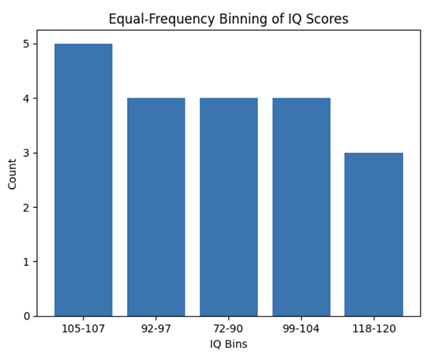


1. Equal-frequency binning using 5 bins

No. of instances / bins

20/5 = 4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | IQ | Bin no |  | ID | IQ | Bin no |
| 18 | 72 | --1 |  | 4 | 101 | --3 |
| 3 | 83 | --1 |  | 20 | 104 | --3 |
| 12 | 88 | --1 |  | 11 | 105 | --4 |
| 14 | 90 | --1 |  | 10 | 106 | --4 |
| 1 | 92 | --2 |  | 13 | 106 | --4 |
| 6 | 92 | --2 |  | 2 | 107 | --4 |
| 9 | 93 | --2 |  | 5 | 107 | --5 |
| 15 | 97 | --2 |  | 16 | 118 | --5 |
| 7 | 99 | --3 |  | 8 | 119 | --5 |
| 19 | 100 | --3 |  | 17 | 120 | --5 |



**Question 4:-**

Chapter 3, exercise 7

1. It seems like you're describing a normal distribution of heights among employees, with a bell-shaped curve centered around 150-175 units. A normal distribution, also known as a Gaussian distribution or bell curve, is a statistical distribution that is symmetric and forms a bell shape when plotted. The characteristics you mentioned, such as the symmetric distribution and the peak at 150-175 units, are consistent with a normal distribution. In a normal distribution:

The majority of data falls within one standard deviation of the mean (average).

The mean, median, and mode are all located at the center of the distribution.

The distribution is symmetric, with tails extending equally in both directions.

If employee heights follow a normal distribution, it can simplify statistical analysis and modeling, as many statistical methods assume a normal distribution. Understanding the distribution of employee heights can be useful for various purposes, such as setting ergonomic standards, designing workspaces, or even predicting health outcomes associated with height.

1. The number of prior criminal convictions held by people given prison demonstrates following qualities: 1. It’s an exponentially distributed histogram. 2. For exponentially distributed histograms the likelihood of low values occurring is very high, but it declines rapidly for higher values, hence likelihood of outliers is high. 3. Hence, we can say that prior criminal convictions held by people are high at the beginning but later they tend to diminish. This data follows an exponential distribution. There is a strong central tendency around 0 and ever decreasing probability of seeing higher values. Data following an exponential distribution can be hard to manage as there tend to be significant outliers, which can upset modelling algorithms. In this case there is at least one instance of a person having over 40 prior convictions, which is very unusual.
2. The LDL cholesterol values for large group of patients demonstrates following qualities: 1. It’s a multimodal distributed histogram. 2. We can observe two groups of patients clearly smokers and non-smokers, it’s quite significant that smokers have peak LDL cholesterol values of about 0.025 whereas non-smokers tend to have lower LDL cholesterol values of about 0.015.
3. The employee ID of university staff demonstrates following qualities: 1. It’s a uniformly distributed histogram. 2. We can see that employee ID are uniformly distributed in all range values.
4. The salaries of motor insurance policy holders demonstrate following qualities: 1. It’s a Unimodal right skewed histogram. 2. It’s evident that a largen number of holders have a salary between 30K to 50K and only a few holders have salary greater than 75k.

**Question 5:-**

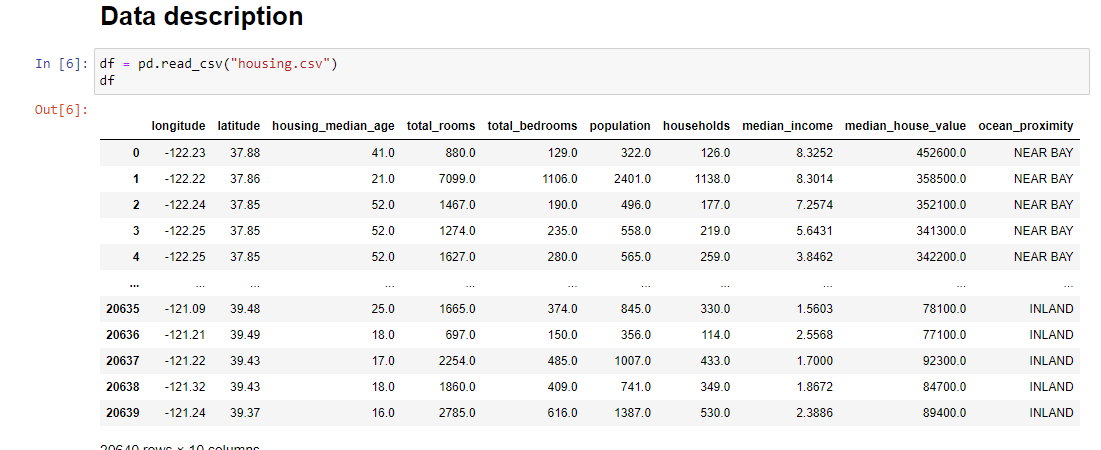
Chapter 3, exercise 9

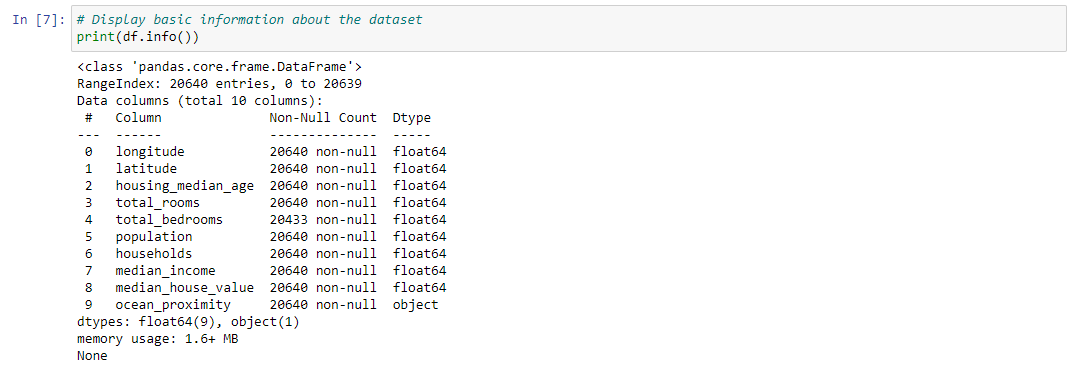
1. The following conclusions on missing values can be made based on the report generated:
   1. There’s 44.02% of missing data of patients who previously suffered tachycardia.
   2. Approximately 2.01% of missing data of patients who suffered tachycardia.
   3. Also, there’s almost 13.03% of missing data of H.R diff i.e., people whose heart rate was different in the current visit and their last visit to the clinic.
2. The following conclusions on irregular cardinality can be made based on the report generated:
3. The gender column has cardinality of 4 meaning along with male and female they have sometimes used m or f to represent gender instead, which is unnecessary and can be standardized.
4. The prev. tachy field is a Boolean field but shows the cardinality of 3 meaning it has some null or incorrect values, Boolean values must only contain 2 values and must be standardized.
5. Similarly, Tachycardia field is a Boolean field but shows the cardinality of 3 meaning it has some null or incorrect values, Boolean values must only contain 2 values and must be standardized.
6. The following conclusions on outliers can be made based on the report generated:
7. The minimum age is 1.00, which seems unusually low. It might be an entry error or a specific case.
8. The minimum weight is 0.00, which is impossible for an adult human. This could be a data entry error or missing data.
9. The minimum BMI is 0.00, which is also unrealistic. BMI should typically not be below 18.5, as it indicates underweight.
10. The maximum value of 596,495.39 is extremely high and likely an error.
11. The maximum value of 77.068.75 is also very high and likely an error.
12. The negative minimum value of -50.00 might indicate an issue or error in data collection. Heart rate difference is typically a positive value, representing changes between resting and active states.
13. The following conclusions on feature distributions can be made based on the report generated:
    1. The distribution of ages appears to be slightly positively skewed. This suggests that most individuals in your dataset are in the higher age range, with relatively few younger individuals.
    2. The distribution of weights appears to have a relatively small standard deviation. This suggests that the weights in your dataset are relatively consistent, with most individuals having similar weights.
    3. The distribution of height suggests that there may be some variation in height within the dataset, but it's not extremely skewed in either direction.
    4. The distribution of BMI values has a wide range. This indicates a considerable variation in BMI values within your dataset, with some individuals having very high or very low BMI values.
    5. The distribution suggests that systolic blood pressure values are somewhat normally distributed, with most individuals falling around the mean.
    6. Similarly, the distribution appears somewhat normal, indicating that diastolic blood pressure values are clustered around the mean.
    7. The distribution of heart rate differences appears to have a negative minimum value of -50.00. A negative heart rate difference is unusual and might indicate measurement errors or inconsistencies in data collection.

**Question 6:-**

**Data Loading and Description:**

****

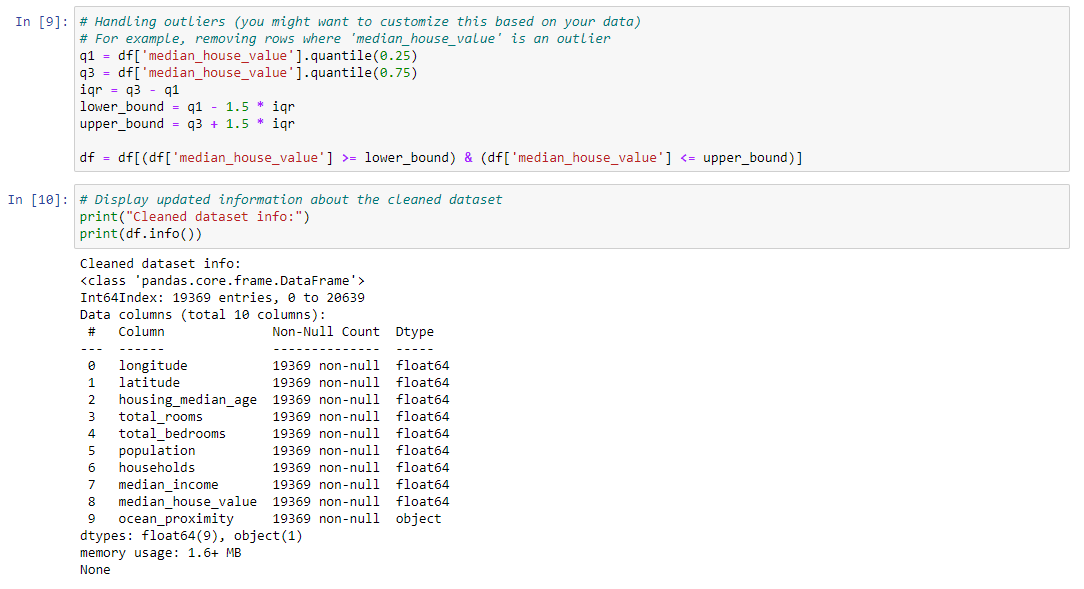
****

****

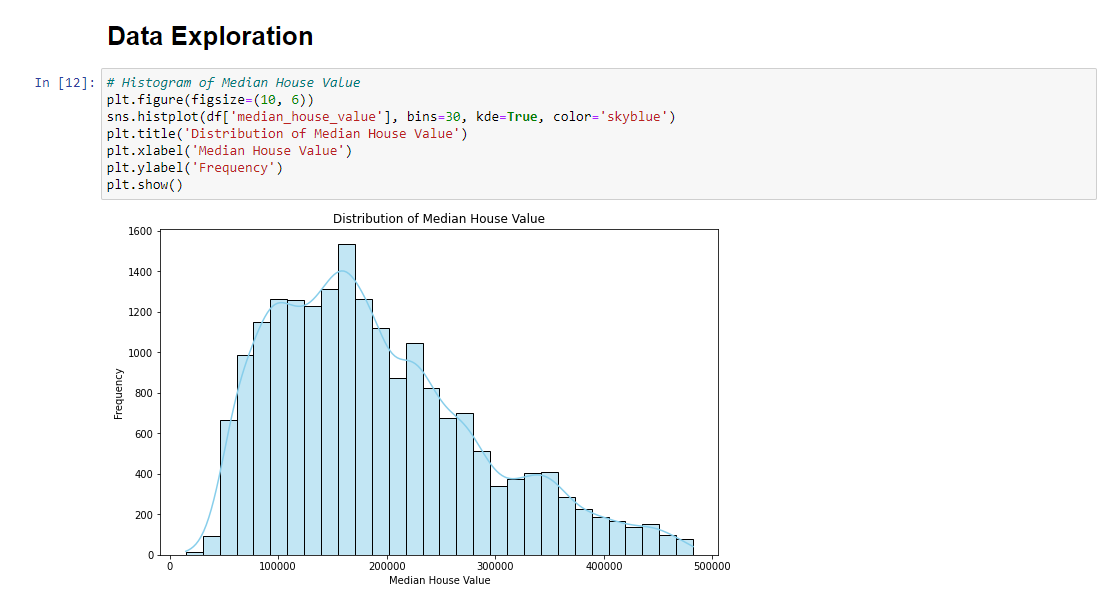
This section loads the dataset into a Pandas DataFrame and prints basic information about the dataset using the info() method. This information includes the number of non-null values, data types, and memory usage.

**Data Cleaning:**

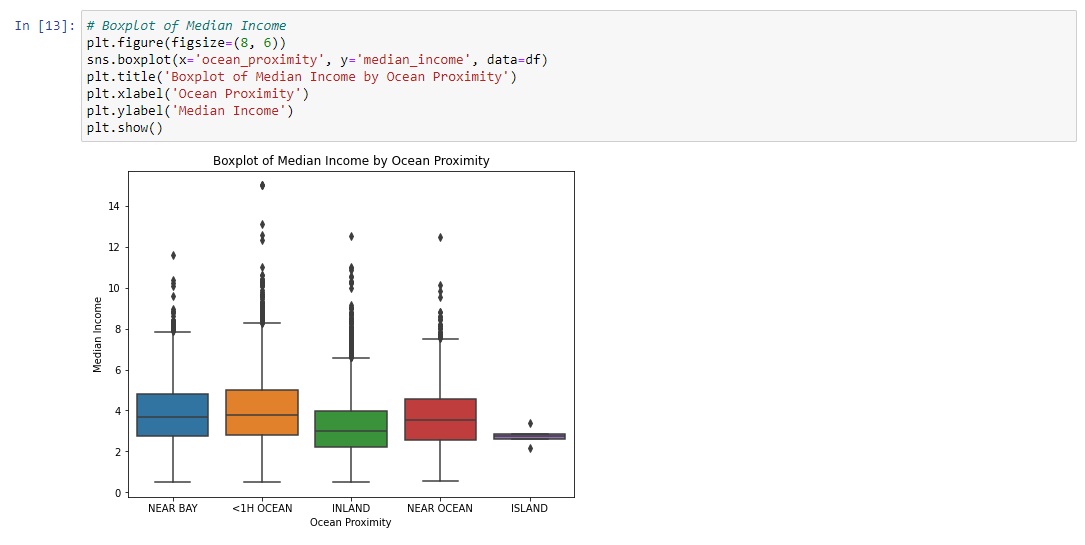
****

****

The code removes rows with missing values using the dropna() method. This is a basic approach to handling missing data, and the inplace=True parameter modifies the DataFrame in place. This section identifies and removes outliers in the 'median\_house\_value' column using the interquartile range (IQR) method. Rows with values outside the defined lower and upper bounds are filtered out. After cleaning, it prints updated information about the cleaned dataset.

****

This section generates a histogram to show the distribution of median house values. The histogram is accompanied by a kernel density estimate (KDE) and is plotted in sky blue.

****

A boxplot is created to display the distribution of median income across different ocean proximities. A pair plot is generated to visualize pairwise relationships between the selected numerical columns. The diagonal shows histograms, and scatter plots are displayed for each pair of variables. These visualizations provide insights into the cleaned dataset, helping to understand the relationships between different variables and the overall distribution of data. Adjustments can be made based on specific analysis goals and dataset characteristics.





