

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: From the EDA on categorical variables, below are the statements for different variables:

- Majority of people seem to have rented / used the bike during fall season, followed by Summer then Winter especially in 2019.
- Customers have increased from 2018 to 2019 significantly.
- Majority of customers are in Sep, August and July.
- During clear weather there are many people riding/using bikes followed by Cloudy. Lesser people are using bikes during light Rain. No customers during Heavy Rain.
- There is a significant increase in customers in 2019 during Clear Weather.
- As temp increases the number of customers for Bike sharing Company increases.
- In 2019, temp was relatively higher, humidity and windspeed has been comparatively lower than previous year. This could be the reason for the increase in the number of people renting/using bikes.
- During holidays, people are renting/using more bikes.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans : `pd.getdummies()` creates a new variable for each level of the category, one column for each unique value of the column and wherever this value is true it is indicated by 1 else 0.

Drop_first=True excludes the dummy variable for the first category of the variable you're operating on. For a P-variables i.e n -levels in a categorical variable, number of dummy variable needed is p-1

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: Temp seems to have highest correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: Below assumptions have been validated after building the model:

- Mean of Residuals:
 - One of the assumptions of linear Regression is that the mean of residuals should be zero.
 - Mean of errors value that I got was $-1.7763568394002505e-15$ which is very close to 0, thus satisfying the assumption
- Homoscedasticity:
 - It means residuals have almost the same variance across the regression line which could be seen in the residual plot.

- From the graph we can say it is unbiased and homoscedastic.
- Normality of error terms/ residuals:
 - Error terms should be normally distributed
 - The density distribution graph depicts a normal distribution curve which satisfies the assumption
- Autocorrelation of residuals
 - There should not be autocorrelation in the data so the error terms should not form any pattern.
 - Performed Durbin-Watson Test and value was 2.073 which states that assumption satisfied.
- No perfect multicollinearity:
 - Heatmap drawn and analysing Pearson's Coefficient suggests that no perfect multicollinearity exists thus satisfying the assumption

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: The top three features which contributes significantly are:

- **Temp**
 - Coefficient Value : 0.490428
- **Yr**
 - Coefficient Value : 0.232943
- **Weathersit_Light_Rain**
 - Coefficient Value : -0.306094

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear Regression is a supervised learning algorithm that uses one or more input variables, also known as independent variables to predict a single dependent output variable. It is used for finding the relationship between two or more variables and predicting results based on the relationship.

The equation creates a line and hence the term linear that best fits the X and Y variables provided which is mathematically represented as:

$$Y = mX + b$$

Where

- Y describes the independent variable which we are trying to predict
- X describes the input variable
- m describes the slope of the line also known as regression coefficient
- b describes the intercept at Y-axis

Based on the signs of regression coefficient, below can be said about linear Regression:

1. A positive sign of regression coefficient shows that as independent variable goes up dependent variable also goes up
2. A negative sign of regression coefficient shows that as independent variables goes down dependent variable go down.

Types:

There are 2 types of Linear Regression:

1. Simple Linear Regression (SLR)
 - a. A simple Linear Regression relies on single variable and its relationship with output variable which is given by below equation:
 - b. $Y = mX + b$
2. Multiple Linear Regression (MLR)
 - a. A multiple or multi-variable linear Regression algorithm determines the relationship between multiple input variables and its relationship with the output variable.

Assumptions:

Below are the assumptions of Linear Regression:

1. Linearity:
 - a. The expected value of a dependent variable is a straight line function of independent variable. The effects of different independent variables are additive to the expected value of dependent variable
2. Statistical Independence:
 - a. The observations are independent of each other
3. Homoscedasticity:
 - a. The variance of errors is a constant across all levels of independent variables
4. Normality:
 - a. The errors follow a normal distribution
5. No multicollinearity:
 - a. The independent variables are not highly correlated with each other.
6. No endogeneity:
 - a. There is no relationship between the errors and independent variables

2. . Explain the Anscombe's quartet in detail.

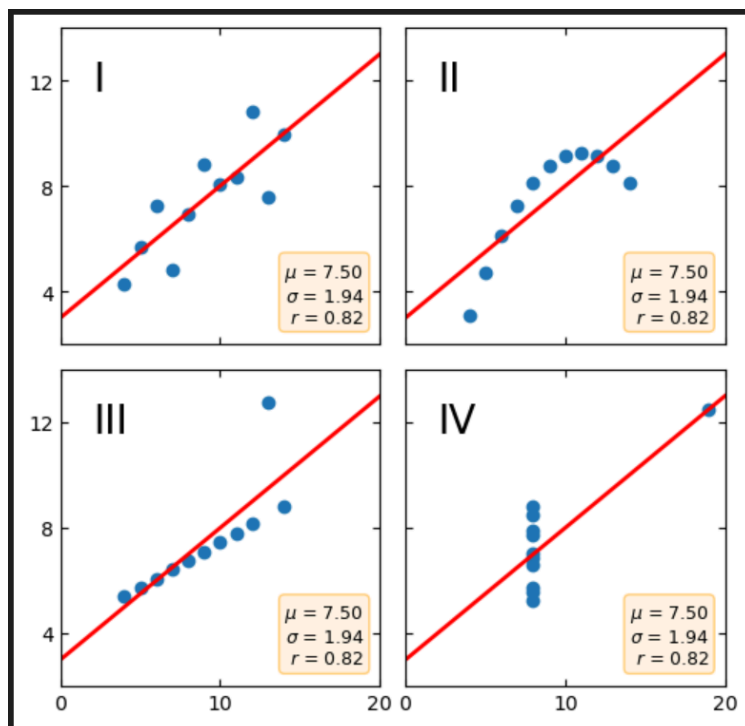
(3 marks)

Ans: Anscombe's quartet was developed by Francis Anscombe in 1973 which is a group of datasets (X, Y) that have the same mean, standard deviation, and regression line, but which are qualitatively different - different graphical representation.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistics properties.

It can be described as a group of 4 datasets which have nearly the same descriptive statistics but have distributions and appear different when plotted on a scatter plot.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



From the given dataset, we notice that mean, std. Deviation and correlation coefficient i.e how strong a relationship is between two variables is same for 4 datasets.

As we can see from scatter plots, each data generates a different kind of plots irrespective of same descriptive statistics.

- Fig. 1 fits the linear regression model pretty well
- Fig 2. cannot fit the linear regression model because the data is non-linear.

- Fig. 3 shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Fig. 4 : shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

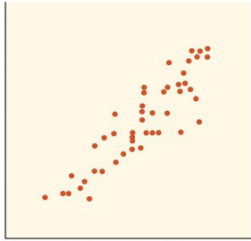
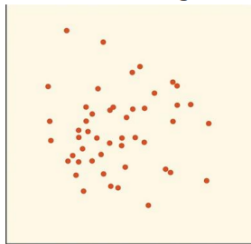
Now, we can say that Anscombe's quartet helps to understand the importance of data visualisation and how easy it is to fool a regression algorithm.

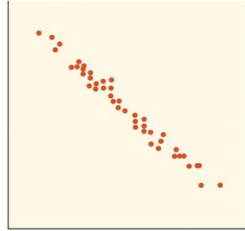
3. What is Pearson's R?

Ans: Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between different variables and their relationships.

The coefficient not only states the presence or absence of correlation between the two variables but also determines the exact extent to which those variables are correlated. It is independent of the unit of measurement of the variables where the values of correlation coefficient can range from +1 to -1.

The correlation coefficient between the variables is symmetric, which means that the value of correlation coefficient between Y and X or X and Y will remain the same.

Pearson correlation coefficient (r)	Correlation type	Interpretation	Graph
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	 <p>Correlation $r = 0.9$</p>
0	No correlation	There is no relationship between the variables.	 <p>Correlation $r = 0$</p>

Pearson correlation coefficient (r)	Correlation type	Interpretation	Graph
Between 0 and -1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	 <p>Correlation $r = -0.99$</p>

To describe a relationship below should be considered:

1. The strength of relationship is given by correlation coefficient
2. The direction of relationship, which can be positive or negative, based on the sign of correlation coefficient
3. The shape of relationship which must always be linear to compute a Pearson correlation coefficient

Whenever a statistical test is conducted between two variables, it is always a good idea to analyse and calculate the value of correlation coefficient to know how strong the variable is.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: It is a technique to standardise the independent variables present in the data in a fixed range. It is performed during data preprocessing to handle highly varying magnitudes or values or units. It also helps in speeding up the calculations in an algorithm.

Why:

Ease of interpretation:

It becomes easy to compare coefficients when they are brought to the same scale. When features are on different scale then interpretation becomes difficult.

Faster Convergence for gradient descent methods:

Usually gradient descent algorithm runs behind the scene. If features are on a different scale, say one feature is between 0-10 and another from 100 - 1000 then it takes a lot of time for convergence. If you bring the features to the same scale then convergence becomes much faster.

Scaling just affects coefficients and not other parameters like t-statistic, F-statistic, p-values etc.

Normalized/ Min -Max Scaling:

This method scales the model using minimum and maximum values.

MinMax Scaling : $X_{\text{scaled}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$

Where

X is original feature

X_scaled is scaled feature

X_min is the minimum value of X

X_max is the maximum value of X.

It brings all of the data in the range of 0 and 1. Use MinMaxScaler from Sklearn to perform scaling.

Standardization:

It basically brings all of the data into a standard normal distribution with mean 0 and standard deviation 1.

Formula $X_{\text{scaled}} = (X - X_{\text{mean}}) / X_{\text{std}}$

Where

X is original feature

X_scaled is scaled feature

X_mean is the mean of X

X_std is the standard deviation of X.

Values on a scale are not constrained to a particular range. This process is called Z-score normalisation. Use StandardScaler from sklearn to perform standardized scaling.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: VIF determines the strength of correlation between independent variables. It represents how well the variable is explained by other independent variables. A common heuristic which is followed is:

VIF : 1 - No multicollinearity

VIF : 4- 5 - Moderate

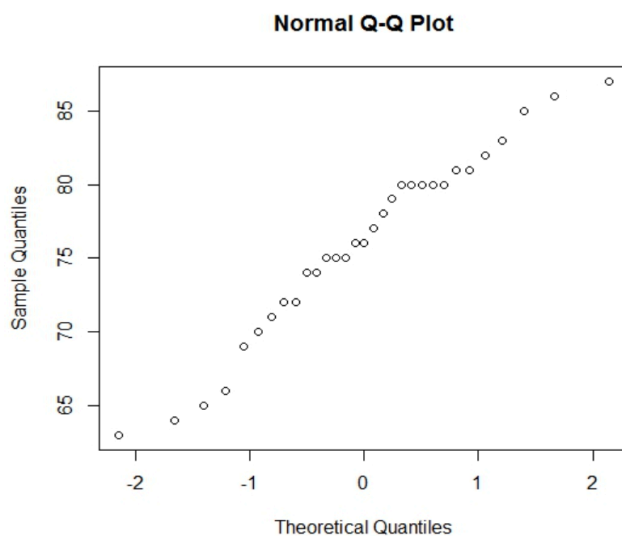
VIF : 10 or greater - Severe

When the VIF value is infinite it shows a perfect correlation between two independent variables. In case of perfect correlation, we get $R^2 = 1$ which leads to infinity for VIF.

To solve the problem, Dropping one of the correlated features will help in bringing down the multicollinearity between correlated features.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans: Q-Q plot also known as quantile-quantile plot is a graphical technique for determining if two datasets come from populations within a common distribution. It is a plot of quantiles of the first dataset against the quantiles of the second dataset. If both sets of quantiles came from the same distribution, we should see the points forming a line thats roughly straight. Below is an example of such a plot:



Advantages:

1. The sample sizes do not need to be equal
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, presence of outliers.

This helps in the scenario of linear regression when we have training and test data set received separately, then we can confirm using Q-Q plot that both datasets are from populations with the same distributions.

It can also used to find if two datasets have common location and scale, have similar distributional shapes and have similar tail behaviour.

This helps in providing a powerful visual assessment, pinpointing deviations between distributions and identifying data points responsible for them.

