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?

ANS:

Here are some of the inferences on the analysis of the categorical variables and their effect on the dependent variable.

1. The season of Fall has the highest median followed by summer as they have the best weather conditions.
2. The median bike rentals have increased in the year 2019 compared to the year 2018. This may be due to the people getting conscious about the environment.
3. The bike rentals are more on non-holiday days compared to holiday. This indicates that people prefer to spend time at home during the holidays.
4. The months of Fall - June to October have a higher median value.
5. The overall median for the weekdays and working-days are the same.
6. The Clear weather situation has the highest median while the weather situation of Light snow has the least.

The count of bike sharing is Zero for the weather situation - 4 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog'.

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

It is important to use drop_first = True as it helps in reducing extra column created during dummy variable creation. It helps to reduce the correlations created among dummy variables.

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

attempt and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables

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

According to the assumption three is a linear relationship between the features and rectilinear regression captures only linear relationship. This can be valid by plotting a scatter plot too

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

The top 3 are temperature weather and year

General Subjective Questions

1) Explain the linear regression algorithm in detail?

ANS :

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product count, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (X) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Equation for linear regression algo:

$y= a0+a1X+ \varepsilon$

Y= Dependent Variable (Target Variable)

x= Independent Variable (predictor Variable)

a0= intercept of the line

a1 = Linear regression coefficient

ε = random error

Types of Linear Regression

1. Simple Linear Regression: If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
2. Multiple Linear regression: If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

---

2) Explain the Anscombe's quartet in detail.

ANS :

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (XXY) points. They were constructed in 1973 by the

statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

---

3) What is Pearson’s R?
ANS:
The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
Between 0 and 1 - Positive correlation (When one variable changes, the other variable changes in the same direction.)
0 - No correlation (There is no relationship between the variables.)
Between 0 and –1 - Negative correlation (When one variable changes, the other variable changes in the opposite direction.)

Pearson correlation coefficient (r) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
ANS:
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
m = (x -xmin) / (xmax -xmin)
Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
z = (x - u) / s

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?
ANS : If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
ANS :
Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.
Normal distributions
We regularly make the assumption of normality in our distribution as we perform statistical analysis and build predictive models. Machine learning algorithms like linear regression and logistic regression perform better where numerical features and targets follow a Gaussian or a uniform distribution.

It’s an important assumption as normal distribution allows us to use the empirical rule of 68 – 95 – 99.7 and analysis where we can predict the percentage of values and how far they will fall from the mean.

In regression models, normality gains significance when it comes to error terms. You want the mean of the error terms to be zero. If the mean of error terms is significantly away from zero, it means that the features we have selected may not actually be having a significant impact on the outcome variable. It’s time to review the feature selection for the model.