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?

Answer:

* Except Spring, All other seasons like Summer, fall and winter seems to experience similar number of total bike riders

* There has been an increase in total bike riders from 2018 to 2019 across all seasons

* Months like May, June, July, August, September, October experience relatively similar and higher number of bike rentals than other months

* There is a “sudden peak in the 50% percentile of bike rentals from March to April” and a “sudden drop in the 50% percentile of bike rentals from March to April”

* On almost all weekdays in 50% of cases roughly ~4500 rentals is expected

* Saturday, wednesday and thursday have larger spread of people(25-75 percentile) taking rentals than other weekdays

* Snowy experiences the least amount of rentals followed by misty days

* Clear weather experience most amount of total rentals in a day

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

Answer:

drop_first is set to True to ensure the number of dummy variables created are n-1 categories and not n categories as If a row has values set 0 set of all the dummy variables would mean it’s belongs to the left out dummy variables category

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

Answer:

Temperature feeling temperature variable have moderately high correlation with the target variable

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

Answer:

Performed validation by performing residual analysis of the training data and plotting the error terms

- Predicted the value of target variable using the training data
- Plotted the value of difference between actual target variable and the predicted target variable
- Ensured that the error terms plotted has uniform distribution and is centred around 0

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

Answer:

Top 3 features:

1. **Temperature**(Higher the temperature higher probability a lot of rentals)
2. **Weather being snowy(**People generally tend to take lesser rentals on Snowy days)
3. **Wind Speed and humidity(**People generally tend to take lesser rentals on days having high wind speed followed by humid days)
4. **Saturdays(**Company can focus on promoting bike rentals on saturday and winter season**)**

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Answer:

Linear regression is a technique used for performing data analysis on continuous variables where-in we try to predict the value of a target/dependent variable based on a set of independent variables. By using this technique we try to find a correlation/linear relationship between target and independent variables which

have a high amount of significance to the model. This technique requires scaling the independent variables and There is generally a constant involved in the equation
Once we have built the model we interpolate/predict the value of target variables for a new set of data.

2.Explain the Anscombe's quartet in detail.
Answer:
Anscombe's quartet can be considered a set of four datasets which come from different distributions yet show descriptive analytics which we get using the describe() function. If we just look at the descriptive analytics that could assume the datasets to belong to the same distributions, however when plotted in a scatter plot it would tell a completely different story.
This quartet makes us realise that all the important features of a dataset must be visualised before we take a decision about the choosing the regression or classification algorithm to create our models

3.What is Pearson's R?
Answer
It is a way or method to measure correlation between two variables, more specifically two continuous variables. It's commonly used in linear regressions.

- Pearson Correlation coefficient when measured generally lies between -1 and 1 and can be categorised into negative, neutral or positive correlation.
- It tries to draw a best fit line between the variable to determine their linear relationship and requires variables
- It assumes that the correlation is determined between continuous variables who have a uniform distribution and does not contain extreme outliers

4.What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?
Answer:
Scaling is a method to normalise/readjust the values of a features/variables in order to make them comparable.
Scaling is one of the essential preprocessing step in model building and is performed to make normalise the values of features to fit them within same range to allow model to properly compare the features
Min-max normalisation or normalised scaling is a process to scale the features within a range of [0,1]. It can be used to fit the values within a range of x,y given x and y are real numbers
Standardisation fits the values of features within a range such that the mean of values for the feature becomes 0. Standardisation is not affected by outliers as compared to normalised scaling

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer:
VIF is a factor that helps us determine correlation between the feature variables. As VIF = 1/(1-R^2), If there exists independent variables which have perfect correlation amongst each other R^2 value would become 1 and hence VIF would become 1/0 ~ infinity.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Q-Q plot or Quantile-Quantile plot is a way to compare different probability distributions by plotting their quantiles against each other.
It is an EDA technique which is used to determine if the 2 datasets that we are comparing come from the population with a common distribution.
The quantiles are plotted against a reference line and if the datasets comes from a common distribution that they should ideally fall with the along the reference line and have lesser deviations from the reference line