Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3marks)

Ans 1. cnt(rides) are max seen in the fall season and minimum in the winter season.

- Also, we can say that rides are demanded max in September and minimum in January.
- There is not much difference in weekdays with cnt(rides)
- cnt(rides) are max demanded in clear weather and minimum in light_snowrain.
- Total rides taken are more in holidays, although maximum values of holiday or non-holiday are the same
- Lastly, more rides taken are in the year 2019 as compared to 2018, which means as the year increases, people get more aware, and usage increases.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans 2. Since one of the columns is generated completely from the others, and hence this extra column is not relevant in model building, and hence it is better to drop that first column always. for example, weathersit has 4 columns, out of which one may be dropped, other 3 can explain 4th one.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans 3. Our target variable is cnt(total rides taken), which has a maximum correlation with temperature and atemp = 63% and then with the year which is 57%

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans 4. In our final model i.e., lr_4, there seems to be very low Multicollinearity between the predictors.

- As we have all the vif values lesser than 5
- Linearity is seen in the model by plotting graphs.
- Homoscedasticity is observed in the train model
- The p-values for all the predictors seem to be significant., i.e. 0.
- f(statistics) of lr_4 is also greater than 1
- Also, r2 and adjusted r2 values are quite good, i.e. 83.4% and 80% respectively.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes? (2 marks)

Ans 5. Demand for shared bikes is maximum depends on the following features-

a) atemp,
b) fall season (September),
c) year

General Subjective Questions

Q1. Explain the linear regression algorithm in detail (4marks)

Ans 1. Linear Regression Algorithm is one of the processes in machine learning, which is used to find the inputs which are not present in data with the help of given input and output variables.

Linear regression is one of the easiest and most popular machine learning algorithms. It is a statistical method used for predictive analysis.

It deals with numerical data, so when a variable is categorical, it is first converted to a numeric variable before moving into linear regression model building.

It assumes a linear relationship between the features(which we are considering) and the target variable. For example, predicting the salary by looking at the qualification or predicting the height by looking at the age.

Q2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Q3. What is Pearson’s R? (3marks)

Ans 3. This term is used in Bivariate Analysis. It is the term used to find the correlation between two variables or to find the linear correlation which measures the strength and direction of relationship between two variables.

Its range lies in between -1 and +1, which can be identified as-

- -1 to 0 = negative or opposite correlation. for eg = speed increases, time taken decreases, for same distance
- 0 = No correlation. for eg intelligence of a person will not depend on his height.
- 0 to +1 = positive and strong correlation. for constant speed, if distance increases, time taken will also increase

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans 4. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Most of the time, the collected data set contains features highly varying in magnitudes, units, and ranges. If scaling is not done then the 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.

Normalized scaling - It brings all of the data in the range of 0 and 1.

Standardized scaling - 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).

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans 5. When there is a high correlation between the variables, multicollinearity is seen and if we don't drop highly correlated variables from the dataset, the calculated vif becomes infinite for variables.

for eg- in our bike sharing dataset, temp and atemp were highly correlated with 99%, and hence we dropped temp, which then stopped vif to lead to high or infinite values

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Ans 6. A quantile-quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs. quantiles from a normally distributed curve.

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check the following scenarios:

If two data sets —come from populations with a common distribution, have a common location and scale have similar distributional shapes, and have similar tail behavior