Assignment-based Subjective Questions

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

The variables “holiday”, “weekday”, “season” and “weathersit” have a significant influence on the dependent variable, in this case “cnt". The demand is generally lower for holidays than non-holidays and towards weekends than on weekdays. Winter and spring seasons see a significantly lower demand than summer and fall, while worsening weather conditions as defined by “weathersit” see an understandable fall in demand. Spring and winter are generally accompanied by bad weather conditions such as light rain, snow and blizzards. “Workingday” does not seem to impact the demand very much.

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

There are 2 reasons one should use drop_first = True during dummy variable creation.

- It makes the dataset lean by not including more variables than necessary. If there are n levels of the categorical variable to be dummified, the nth level can be defined as (n-1) dummy columns being equal to zero. So keeping the nth column does not add any value, instead makes the dataset wider.
- Since the nth dummy column is dependent on the other (n-1) dummy columns by means of a NAND function (or a linear combination that can approximate a NAND gate), it violates the assumption of linear independence of predictors in the linear regression model.

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

The highest correlated variables are “temp” and “atemp”. Both of these convey similar information, but “temp” is more business oriented since it is the measured temperature.

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

The validation of assumptions were conducted in 3 steps:

- Normalcy of the residuals distribution was tested with a histogram and a Q-Q plot.
- Heteroscedasticity of the residuals were tested against categorical variables by ANOVA.
- Heteroscedasticity of residuals were tested against the continuous predictors by scatterplot.

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

The top 3 contributing features in order of absolute value of coefficients are:

- "temp"
- "lightrain" (derived from "weathersit", indicates presence of category 3).
- "2019" (indicator of whether or not the year is 2018 or 2019, indicates YoY growth)

General Subjective Questions

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

Linear Regression is a parametric supervised learning algorithm that aims to predict the value of a continuous target variable based on the values of discrete or continuous predictor variables through minimizing the sum of the squares of the residual terms, i.e. the differences between actual Y (target variable) and predicted Y.

If we assume the feature vector to be X, then the equation for linear regression is given as:
$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$,
where β_0 is known as the intercept or the "bias term" and $\beta_1, \ldots \beta_n$ are known as "coefficients" for the features $X_1, \ldots X_n$. The terms $\beta_0, \ldots \beta_n$ are also called "model parameters". The model learns the set of parameters creating a hyperplane that best fits the data.

The learning process can be done by two ways:

- Minimizing the residual sum of squares, i.e. $\sum_i (Y_i - \hat{Y}_i)^2$ by means of differential calculus.
- Using an computationally efficient algorithm called "gradient descent" that starts with an initial value for the β vector and keeps iterating until the best value is reached. This method is implemented in most linear regression packages.

The linear regression algorithm makes the following assumptions:

- X and Y are linearly correlated.
- The features X are independent.
- The residual distributions for each Y_i are independent.
- The residual distributions are identically normal with mean zero and constant variance.

These assumptions need to be validated in order to test whether linear regression is indeed a suitable algorithm for the problem being solved.

Post training, the accuracy of the model is tested using several metrics. Two of the most popular are the R-squared and adjusted R-squared scores, both of which indicate the proportion of variability in the dependent variable that is explained by the independent variables through a linear combination. Adjusted R-squared penalizes an increasing number of features, so it prevents us from adding features that do not explain any additional variability.

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

Anscombe's Quartet is a set of four datasets developed by the scientist Francis Anscombe to demonstrate the importance of visualizing data over numerical summaries. The datasets have the following characteristics:

- They are all of the form (X, Y), i.e. two columns.
- All four datasets have the exact same summary statistics up to a reasonable precision.
- All four datasets show vastly different scatter plots when plotted.

This underlines the importance of creating visualizations to obtain an idea of the patterns alongside numerical summaries instead of simply relying on the summaries themselves.

What is Pearson's R? (3 marks)

Pearson's R is a statistic that measures linear correlation between two continuous variables. It varies between -1 and +1, with values close to -1 indicating a strong negative correlation and values close to +1 indicating a strong positive correlation. Values at or near zero indicate a very weak or no correlation between the variables. The formula is provided as follows:

$$R = \frac{Cov(X,Y)}{Var(X)Var(Y)}$$

where $Cov(X, Y)$ indicates the covariance between X and Y and $Var(X)$ and $Var(Y)$ are the variances of X and Y respectively.

There are certain limitations to using Pearson's R such as:

- It does not effectively capture correlation that is non-linear in nature.
- It assumes all data points of (X, Y) are independent.
- It is sensitive to outliers.

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

Scaling is the process of converting the data to a different scale than originally provided. In data science, scaling is often performed to bring the dataset to a uniform scale. This is done for the following reasons:

- If one or more columns are having vastly different scales than the rest of the columns, then it becomes difficult to visualize the data.
- When predicting the data using parametric techniques, often interpretability becomes challenging because parameters vary by the scale of the variable and higher priority predictors can get lower parameters due to being of a smaller scale.
- Bringing the training data to a uniform scale results in faster convergence for algorithms like gradient descent.

The two most popular methods of scaling in data science are Normalization and Standardization. Normalization, also called Min-Max scaling, compresses the data into 0 and 1, with no data point exceeding either limits. It is done using the formula:

$$\frac{X-X_{min}}{X_{max}-X_{min}}$$

Standardization on the other hand rescales the data between -1 and 1 with mean at 0. The formula is provided as:

$$\frac{X-\bar{X}}{\sigma(X)}$$

Often normalization is the preferred approach for the following reasons:

- It can handle outliers in the way that it squeezes them within the limits. Extreme values occupy values very close to 0 or 1 after rescaling but still remain within the limits. Standardization is able to center the mean at 0 but does not do a good job at keeping outliers or extreme values within -1 and 1.
- Standardization assumes data to follow a normal distribution, which may not be a good assumption for a small number of data points.

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

The formula for VIF is given by

$\frac{1}{1-R_i^2}$, where R_i^2 is the coefficient of determination or r-squared score for the ith predictor. For VIF to be infinite, therefore, the denominator needs to be 0, which means $R_i^2 = 1$. This would imply that 100% of the variance of the i-th predictor is explained by the other predictors; in other words, the i-th predictor is a linear combination of the remaining predictors.

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

A Q-Q plot is a scatterplot between quantiles of a data series X and theoretical quantiles obtained from a normal distribution. If the distribution of X is normal or close to normal, the scatterplot would resemble a highly linear correlation between the two sets of quantiles. On the other hand, if it is indicative of a weak correlation, this implies that X is not normally distributed nor can be approximated well by a normal curve.

In linear regression, a Q-Q plot can be used to test the normalcy of the residual distribution. One of the assumptions of a linear regression model is that the distribution of residuals has to be normal, so the Q-Q plot provides an excellent visual to verify this. Q-Q plot, alongside a histogram showing the shape of the residual distribution, can be a highly accurate way to tell whether the assumption is validated or not.