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. The dataset contained six categorical variables. We studied their effect on the dependent variable ('cnt') using a Box plot (see figure above). We could draw the following conclusions: season: Season 3 accounted for about 32% of all bike bookings, with a median of over 5000 bookings (during a two-year period). Seasons 2 and 4 followed, accounting for 27% and 25% of total bookings, respectively.

This suggests that the season can be an excellent predictor of the dependent variable. mnth: Almost 10% of bike bookings occurred in the months 5, 6, 7, 8, and 9, with a monthly average of around 4000 bookings. This suggests that mnth has a strong trend and can be a good predictor of the dependent variable. Nearly 67% of bike bookings occurred during 'weathersit1,' with a median of close to 5000 bookings (during a two-year period).

Weathersit2 came in second with 30% of total bookings. This suggests that whether or not there is a trend towards bike bookings, it can be a good predictor of the dependent variable. holiday: Almost 97.6% of bike bookings occurred while it was not a holiday, indicating that this data is definitely skewed. This means that the holiday cannot be used to forecast the dependent variable. weekday: The weekday variable indicates a relatively close trend (between 13.5%-14.8% of total bookings on all days of the week), with independent medians ranging from 4000 to 5000 bookings.

This variable may have little or no effect on the prediction. We'll let the model decide whether or not this should be included. workingday: About 69% of bike bookings occurred during the 'workingday,' with a median of close to 5000 bookings (during a two-year period). This suggests that the workingday can be a good predictor of the dependent variable.

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

Ans. drop_first=True is useful because it reduces the unnecessary column created during dummy variable generation. As a result, it lowers the correlations formed between dummy variables. Assume we have three different types of values in the Categorical column and want to construct a dummy variable for that column.

If one variable is neither furnished nor semi_furnished, it is clearly unfurnished. As a result, we don't require a third variable to identify the unfurnished. As a result, if we have a categorical variable with n levels, we must utilise n-1 columns to represent the dummy variables.

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

Ans. The relationship between temp and atemp is linear. Due to multicolinearity, neither of the parameters can be employed in the model. Based on VIF and p-value in relation to other variables, we will pick which parameters to preserve.

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

Ans. The predictors have very low multicollinearity, and the p-values for all of them appear to be significant. For the time being, we will consider this to be our final model (unless the Test data metrics are drastically different).

The model's coefficient values for all variables are not equal to zero, indicating that we can reject the model. The Null Hypothesis F-Statistics is used to test the Model's overall significance: The more the F-Statistics, the more significant the Model. Prob (F-statistic): 3.77e-181 F-statistic: 233.8 The whole model is significant, as indicated by the F-Statistics value of 233 (which is more than 1) and the p-value of '0.0000'.

After plotting the histogram, the residuals were found to be regularly distributed. As a result of our valid assumption for Linear Regression, we can see that there is no multicollinearity between the predictor variables, as all of the values are within the allowable range of less than 5.

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

Ans. According to our final Model, the top three predictor variables that influence bike booking are: Temperature (temp) - A coefficient value of '0.5636' showed that increasing the temp variable by one unit increases the number of bike hires by 0.5636 units. Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3070' indicates that a unit rise in the Weathersit3 variable reduces the number of bike hires by 0.3070 units in comparison to Weathersit1.

Year (yr) - A coefficient value of '0.2308' showed that increasing the yr variable by one unit increases the number of bike hires by 0.2308 units. As a result, it is advised to prioritise these variables when planning in order to achieve maximum Booking. The following best aspects to examine are season_4: - A coefficient value of '0.128744' indicates that a unit increase in the season_4 variable increases the bike hire numbers by 0.128744 units in relation to season_1. windspeed: A coefficient value of '-0.155191' showed that a unit increase in windspeed variable reduces the number of bike hires by 0.155191 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression is a supervised learning-based machine learning technique. It runs a regression test. Based on independent variables, regression models a goal prediction value. It is mostly used to determine the link between variables and predicting. Different regression models differ in the type of relationship they evaluate between dependent and independent variables, as well as the amount of independent variables utilised.

Linear regression is used to predict the value of a dependent variable (y) based on a given independent variable (x). As a result, this regression technique establishes a linear relationship between x (input) and y (output). As a result, the name Linear Regression was chosen. In the diagram above, X (input) represents job experience and Y (output) represents a person's wage.

Our model's best fit line is the regression line. y = a1 +a2.Here, x is intercepted by a1. a2 is the x x: input training data coefficient. y: data labels We get the best fit line after we get the best 1 and 2 values. So, when we use our model to forecast, it will predict the value of y for the input.

2. Explain the Anscombe’s quartet in detail

Ans. Combe's Quartet is a collection of four data sets that are virtually identical in simple descriptive statistics, but have some idiosyncrasies that deceive the regression model if formed. They have highly diverse distributions and show very differently on scatter plots. Francis Anscombe to demonstrate the significance of plotting graphs before analysing and modelling, as well as the impact of other observations on statistical features.

There are four data set plots that contain virtually identical statistical observations and provide the same statistical information, which includes the variance and mean of all x,y points in all four datasets. This emphasises the importance of visualising the data before applying various algorithms to build models out of it, implying that the data features must be plotted in order to see the distribution of the samples, which can help you identify the various anomalies present in the data such as outliers, diversity of the data, linear separability of the data, and so on.

The first of the four datasets is Dataset 1: it fits the linear regression model quite well. Dataset 2: Because the data is non-linear, a linear regression model could not be fit quite properly. Dataset 3: displays the dataset's outliers that cannot be handled by the linear regression model. Dataset 4: displays the dataset's outliers that cannot be handled by the linear regression model.

3. What is Pearson’s R?

Ans. Pearson's r is a numerical representation of the strength of the linear relationship between the variables. The correlation coefficient will be positive if the variables tend to rise and fall together. The correlation coefficient will be negative if the variables tend to go up and down in opposite directions, with low values of one variable correlated with high values of the other.

As the scatterplot of weight against height for a sample of older women indicates, "tends to" suggests the link holds "on average," not for every arbitrary set of observations. Height and weight tend to rise and fall together, hence the correlation coefficient is positive. However, as the points in the two boxes show, it is possible to identify pairs of people in which the taller individual weighs less.

Pearson's correlation coefficient ranges between -1 and +1 in the following cases: r = 1 indicates that the data is perfectly linear with a positive slope (both variables tend to vary in the same direction). r = -1 indicates that the data is perfectly linear with a negative slope (both variables tend to vary in opposite directions). r = 0 indicates that there is no linear relationship. r > 0 5 indicates a weak connection. r > 5 8 indicates a moderate relationship. r > 8 indicates a strong relationship. The graph below depicts several data sets and their correlation coefficients. The first data set has a correlation coefficient of 0.996, the second has a correlation coefficient of -0.999, and the third has a correlation coefficient of -0.233.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardised scaling?

Ans. Scaling is a data Pre-Processing step that is used to independent variables in order to normalise the data within a specific range. It also aids in the speeding up of algorithm calculations. Most of the time, the acquired data set contains features with widely disparate magnitudes, units, and ranges. If scaling is not performed, the method simply considers magnitude rather than units, resulting in erroneous modelling.

To solve this problem, we must scale all of the variables to the same magnitude level. It is vital to note that scaling has no effect on the other parameters such as t-statistic, F-statistic, p-values, R-squared, and so on. 1- Normalization/Min-Max Scaling: This reduces all data to values between 0 and 1. sklearn.preprocessing.MinMaxScaler aids in the implementation of normalisation in Python. 2- Scaling by standardisation: Standardisation replaces values with their Z scores.

It transforms the data into a conventional normal distribution with a mean () of zero and a standard deviation of one (). sklearn.preprocessing.Python's scalability aids in the implementation of standardisation. One disadvantage of normalisation over standardisation is that it removes some data information, particularly about outliers.

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

Ans. The variance inflation factor (VIF) measures the degree of correlation between one predictor and the others in a model. It is used to detect collinearity and multicollinearity. Higher values indicate that precisely assessing the contribution of predictors to a model is difficult to impossible. VIF = 1/1-R^2 VIF = infinity if there is perfect correlation. A high VIF number suggests that there is a relationship between the variables.

If the VIF is 4, it signifies that multicollinearity has inflated the variance of the model coefficient by a factor of four. This means that the standard error of this coefficient has been increased by a factor of two. The confidence interval of the model coefficients is determined by the standard error of the coefficient. Because of the presence of multicollinearity, if the standard error is big, the confidence intervals may be enormous, and the model coefficient may be non-significant.

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

Ans. Q Q Plots (Quantile-Quantile Plots) are comparisons of two quantiles. A quantile is a fraction of values that fall below that quantile. The median, for example, is a quantile where 50% of the data falls below it and 50% fall above it. The goal of Q Q plots is to determine whether two sets of data are from the same distribution. On the Q Q plot, a 45-degree angle is drawn; if the two data sets are from the same distribution, the points will fall on that reference line. The quantile-quantile (q-q) plot is a graphical tool for detecting whether two data sets are from the same population.

A q-q plot is a comparison of the first data set's quantiles to the quantiles of the second data set. The slope tells us whether our data's steps are too large or too tiny. If we have N observations, for example, each step traverses 1/(N-1) of the data. So we're looking at how the step sizes (also known as quantiles) in our data compare to the normal distribution. A steeply sloping section of the QQ plot indicates that the observations in this part of our data are more spread out than we would expect if they were normally distributed. One possible explanation for this would be an extremely high number of outliers (as in the QQ plot we created). An exceptionally large number of outliers (as in the QQ plot we created previously with our technique) could be one explanation of this.