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:

Categorical variables require special attention in regression analysis because, unlike continuous variables, they cannot by entered into the regression equation just as they are. Instead, they need to be recoded into a series of variables which can then be entered into the regression model. There are a variety of coding systems that can be used when recoding categorical variables. Regardless of the coding system you choose, the overall effect of the categorical variable will remain the same. Ideally, you would choose a coding system that reflects the comparisons that you want to make. For example, you may want to compare each level of the categorical variable to the lowest level. In that case you would use a system called **simple coding**. Or Compares each level of a variable to the omitted (reference) level **dummy coding**. By deliberately choosing a coding system, you can obtain comparisons that are most meaningful for testing your hypotheses.

Because dummy coding compares the mean of the dependent variable for each level of the categorical variable to the mean of the dependent variable at for the reference group, it makes sense with a nominal variable.

From our assignment

Summer and Fall season have high bike rental count.

Bike rental count is increasing from Jan to Jun, remains constant till sept and then decreasing.

People rent more bikes in 2019.

people rent less bike during holiday.

In fall season people rent more bike during holiday and in Summer, Winter and Spring less during holiday.

In Summer season people rent more on not working day, while in spring, fall and in winter people rent more bike during working day.

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

Ans:

we encode each category with a different binary feature. In statistics, it is common to encode a categorical feature with k different possible values into k–1 feature (the last one is represented as all zeros). This is done to simplify the analysis (more technically, this will avoid making the data matrix rank-deficient).

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

Ans: temperature or atemp have highest co relation with count [target variable]

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

Ans: we validate the assumption using residual analysis which plot a histogram of error terms, and error should normally distributed. We plot the graph between y_train with y_train_pred.

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

Ans:

The top feature contributing significantly are temperature, windspeed, year also summer season have a significant role too.

General Subjective Questions

1) Explain the linear regression algorithm in detail?

Ans:

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

$Y = \beta_1 + \beta_0 . X$

While training the model we are given:

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
β_1**:** intercept
β_0**:** coefficient of x

Once we find the best β_1 and β_0 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

The following are some assumptions about dataset that is made by Linear Regression model –
Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.

Algorithm:
1. Understand the data set and visualize the patterns, identify target variables and independent variables.
2. Find the correlation b/w all the continuous variable with target variable.
3. Perform segmented univariate analysis on all the categorical variables and try to understand their relationship with target variable.
4. Remove null values if there are any or fill it with mean or median.
5. Create n-1 dummy variables for categorical columns and convert all the columns to numerical values.
6. Split the dataset in train and test.
7. Apply scaling (min-max or standardization) on train set.
8. Create X_train , y_train from train dataset.
9. For Simple linear regression we can directly create model here with one independent variable and check the summary
10. Else for Multiple regression we have 2 options manual backward elimination approach

where we have to take all the features and then remove each one by one after checking p-value and VIF or Recursive feature elimination.

11. If variables are greater than 10 then we can use RFE, where we have to pass the number of top predictors we have to find.
12. Once we got top predictors from step 11 , Create model and check its summary
13. Check the p-value and VIF and remove the features having high p-value (>0.05) also high VIF (>5). Make sure to use below priority while removing features.
• High p-value and high VIF
• High p-value and low VIF

• Low p value and high VIF

14. Continue step 13 until we got all significant features (p-value<0.05) and VIF<5. Make sure to have good R-Square and adjusted r -squared and F statistic (close to zero means overall model is fit)

15. Make predictions which is y_train_pred values and perform residual analysis on final model.

16. Plot the residuals to check normal distribution and constant variance.

17. Scale test set, make sure only to transform and not fit

18. Make prediction on test set.

19. For model comparison check the Adjusted r squared of test set.

20. If there is less than 5 % difference b/w R-Squared and Adjusted r squared of test and train set then it means it's a good model

21. Sort the drivers by coefficients to provide the result to management team to take business strategic decisions.

2) Explain the Anscombe's quartet in detail.

Ans:

Anscombe's Quartet was established by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must highlight **COMPLETELY,** when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

3) What is Pearson's R?

Ans:

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, temp and count. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:
Scaling is required to transform the data set in precise scale either 0-1 or 0-100.
scaling doesn't affect your model. It is awfully important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So, it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:
The most popular methods for scaling:

1) Min-Max Scaling – Normalization

1) It converts all the values in 0 and 1.
2) x-min(x)/max(x)- min(x)

2) Standard Scaling – standardization

   1) With this all the data will be having mean of 0 and standard deviation of 1
   2) X-mean(x)/sd(x)

The difference is that, in scaling, you're changing the range of your data while in normalization you're changing the shape of the distribution of your data.
why scale?

1) Helps with interpretation.
2) Faster convergence of gradient descent

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

Ans:
A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
1. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).
2. The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.
3. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

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

Ans:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

1) It can be used with sample sizes also

2) 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 following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior