Name – Mitali Seth

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 – These are some analysis between categorical and dependent variable

- bike count is more in season 3 (fall)
- bike count is more in 2019
- bike count is more in month from May - Oct
- bike count is more when no holiday
- bike count is more when no holiday or weekends
- bike count is more in clear weather
- There is no significant change in bike demand with workign day and non working day

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

Ans – We use drop_first=True because, if categorical variable has n level build we need only n-1 dummy variables.

Ex: For relation status -> single, married and In Relationship, we need only 2 dummy variables (we can drop single) b/c if In relationship is 0 and Married is 0 this means that person is single.

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

Ans – variable **temp** has highest correlation with target variable.

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

Ans – We can do Residual analysis, to validate final model, if bar plot of residual is normally distributed means our assumptions are valid.

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

Ans – these are the feature contributing significantly :

- Temp = 0.548
- Weathersit - Light Snow = -0.250
- Yr = 0.230

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans - linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.

Equation of linear regression is :

y = b1X1 + b2X2 + b3X3 + …… bnXn + b0

where y is dependent variable X1,X2….. Xn are the explanatory variables. b1, b2, b3….bn are coefficients which explain correlation of explanatory variables with dependent variables and b0 is intercept (constant).

Objective of Linear regression is to find such values of b0 and bi so that sum of square difference of predicted and actual values (RSS) are minimized.

In python we can use OLS() function to get value of b0 and bi.

To minimize cost function (RSS) we can use:

- Differentiation
- Gradient Descent Approach

In Gradient Descent Approach, we calculate R2, which tells how much variance in the data has been explained by the model.

Higher R2 good the model.

Assumptions in LR:

- X and y should display some sort of linear relationship.
- Residual should be normally distributed.
- Residuals error should show constant variance (homoscedasticity).
- Observations are assumed to be independent of each other.

2. Explain the Anscombe's quartet in detail.

Ans - Anscombe's quartet tells that visual exploration of data is crucial. It cautions against relying solely on summary statistics, as different datasets can have similar statistics but tell very different stories when visually examined.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear quite distinct when graphed.

Datasets:

1. Dataset I:
   - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
   - y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82
2. Dataset II:
   - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
   - y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26
3. Dataset III:
   - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
   - y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42
4. Dataset IV:
   - x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8
   - y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91

Descriptive Statistics:

For each dataset, the mean of x, mean of y, variance of x, variance of y, correlation between x and y, and the linear regression line (fitting y = mx + b) are nearly identical.

Implications:

1. Visual Inspection:
   - When you graph these datasets, you see that they have very different patterns.
   - Dataset I: Roughly linear.
   - Dataset II: Non-linear, but still well-fitted by a linear regression.
   - Dataset III: Exhibits a strong outlier that influences the linear fit.
   - Dataset IV: Non-linear, influenced by an extreme outlier.
2. Importance of Visualization:
   - Anscombe's quartet illustrates the importance of visualizing data to understand its underlying patterns.
   - Relying solely on summary statistics can be misleading if the data distribution is not considered.
3. Statistical Caution:
   - Descriptive statistics can be the same for very different datasets.
   - Outliers can heavily influence the results, especially in smaller datasets.

3. What is Pearson’s R?

Ans - Pearson's correlation coefficient, often denoted by the symbol �*r*, is a measure of the strength and direction of a linear relationship between two variables. It quantifies how well the relationship between two variables can be described by a straight line. The coefficient ranges from -1 to 1,

r=1, means perfect +ve relationship between variables
r=-1, means perfect -ve relationship between variables
r=0, means no Linear Relationship

Assumptions for Person’s R:

- Variables should be normally distributed.
- Variables should be linearly distributed.
- There should be no outliers.

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

Ans – In Multiple linear regression we have multiple variables, there can be case that values for some variables can be very high (ex-Price) as compare to other variables, so coefficients value Price will be high and for other it will be low. Hence to avoid model with very weird coefficients that might be difficult to interpret we need to do scaling.

There are two methods for scaling :

1. Standardizing - The variables are scaled in such a way that their mean is zero and standard deviation is one.

   Xi = Xi – X(mean) / X(std)

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between 0 and 1 using the maximum and the minimum values in the data.

Xi = Xi – X(min) / X(max) – X(min)

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

Ans - VIF = 1/ (1 – R2)

So in the case where R2 = 1, i.e it can be case of perfect multicollinearity where one predictor variable is perfect linear combination of other predictor variable, in that case we can see VIF = Inf.

To avoid this so that we don't get unstable coefficients value we need to handle multicollinearity by eliminating that variable which has perfect correlation on other predictor variable.

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

Ans – Q-Q plot helps to find if data is normally distributed or not. It is a plot where y axis is data quantile and x axis is theoretical quantile (ex-theoretical normal distribution, uniform distribution). In the plot we plot the calculated theoretical quantiles against the sorted observed data and draw a straight line which passes from origin with slope 1, if maximum data points lie on this reference line means data is normally distributed.

Importance :

- Normality Assumption – In linear regression residuals should be normally distributed, hence to very that Q-Q plot can help.
- Model Adequacy : If residuals follow normal distribution means that our model is appropriate for data.
- Outlier Detection – If points drawn on Q-Q plot divert very much from reference line indicates there is outliers in our data set.
- It can be used to test distribution amongst 2 different datasets to compare the distributions of these datasets to see if they are indeed the same.