ML C63 - Tejas Kapasi (Kapasitejas@gmail.com)

Linear Regression Subjective Questions

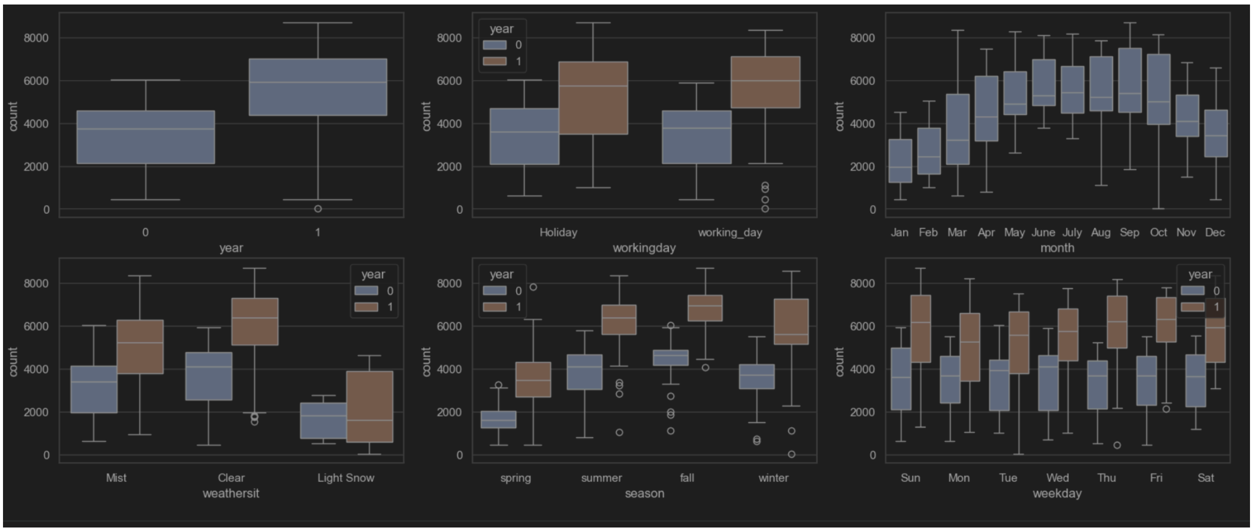
Assignment based Subjective Questions:

1. **Question: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer:

For Categorical variable analysis I have added Box Plots and Bar Plots. Below are the points we can infer:

- **Yearly Comparison:** 2019 saw a greater number of bookings compared to 2018, indicating good business progress.
- **Day of the Week:** Bookings are higher on Thursdays, Fridays, Saturdays, and Sundays compared to the beginning of the week.
- **Weather Impact:** Clear weather conditions have naturally attracted more bookings.
- **Workdays vs. Non-Workdays:** Bookings are almost equal on working days and non-working days.
- **Peak Months:** The majority of bookings occurred during May, June, July, August, and September. The trend increased from the start of the year until mid-year, then began to decline towards the year's end.
- **Seasonal Trends:** The fall season has seen a significant increase in bookings, with each season showing a drastic rise from 2018 to 2019.



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

Answer:

Using drop_first=True is crucial when creating dummy variables, as it helps eliminate the **extra column** that is typically generated.

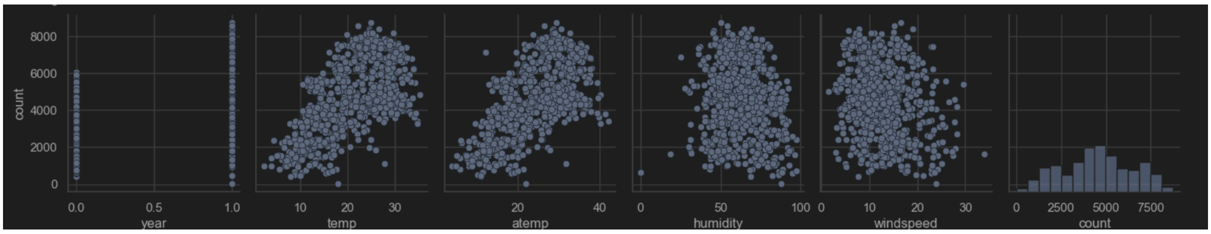
1. **Avoiding Multicollinearity:** When you create dummy variables for a categorical feature, you end up with multiple binary columns. If you include all these columns in your model, it can lead to **multicollinearity**, where one variable can be perfectly predicted from the others. By dropping the first dummy variable, we can avoid this issue.

2. **Simplifying Interpretation:** Dropping the first dummy variable makes the interpretation of the model coefficients easier. The coefficients of the remaining dummy variables represent the difference in the outcome variable compared to the dropped category (the baseline).

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

Answer:

‘Temp’ has the highest correlation with Count Variable.



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

   Answer:

1. Linearity: The relationship between the independent and dependent variables should be linear.
2. Independence: The residuals should be independent.
3. Homoscedasticity: The residuals should have constant variance.
4. No Multicollinearity: Independent variables should not be too highly correlated with each other.
   a. Calculate the Variance Inflation Factor (VIF) for each predictor. A VIF value greater than 10 indicates high multicollinearity.
5. Normality: The residuals (Error Terms) should be approximately normally distributed.

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

   Answer:

General Subjective Questions

1. **Question: Explain the linear regression algorithm in detail.**

   Answer:

In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.
In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

The assumptions of linear regression are:

1. The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'.
2. Assumptions about the residuals:
   a. Normality assumption: It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.
   b. Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
   c. Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ^2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
   d. Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

3. Assumptions about the estimators:
    a. The independent variables are measured without error.
    b. The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

To see if linear regression is suitable for any given data, a scatter plot can be used. If the relationship looks linear, we can go for a linear model. But if it is not the case, we have to apply some transformations to make the relationship linear. Plotting the scatter plots is easy in case of simple or univariate linear regression. But in the case of multivariate linear regression, two-dimensional pairwise scatter plots, rotating plots, and dynamic graphs can be plotted.

A linear regression model is quite easy to interpret. The model is of the following form:

$y=\beta0+\beta1X1+\beta2X2+...+\beta nXn$

The significance of this model lies in the fact that one can easily interpret and understand the marginal changes and their consequences. For example, if the value of x_0 increases by 1 unit, keeping other variables constant, the total increase in the value of y will be β_i. Mathematically, the intercept term (β_0) is the response when all the predictor terms are set to zero or not considered.

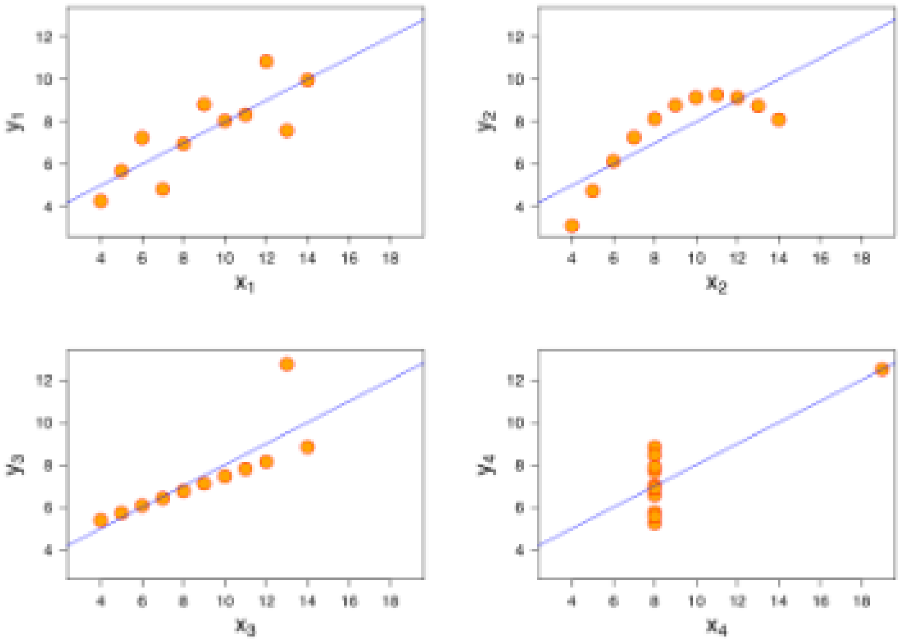
2. Question: Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and the effect of outliers and the distribution of data on statistical properties.

Simple linear regression has quite a few shortcomings:

- It is sensitive to outliers
- It models the linear relationships only
- A few assumptions are required to make the inference

These phenomena can be best explained by the Anscombe's Quartet, shown below:



As we can see, all the four linear regressions are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have gotten a great line fitted through the data points. So we should never ever run a regression without having a good look at our data.

3. Question: What is Pearson's R?

Answer: It is a measure of the strength and direction of the linear relationship between two variables.

Range: It ranges from -1 to 1.

- **1 means a perfect positive linear relationship (as one variable increases, the other also increases).**
- **-1 means a perfect negative linear relationship (as one variable increases, the other decreases).**
- **0 means no linear relationship.**

Interpretation:

- **Positive values (e.g., 0.5) indicate a positive relationship.**

- **Negative values (e.g., -0.5) indicate a negative relationship.**
- **Closer to 0 means weaker relationship.**
- **Closer to 1 or -1 means stronger relationships.**

In essence, Pearson's R helps us understand how two variables move together.

if you have some sense of the relationship being non-linear, you should look at Spearman's R instead of Pearson's R.

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

Answer:

Scaling is a data preprocessing technique used to adjust the range of features in your dataset. It ensures that all features contribute equally to the model's performance by bringing them to a common scale.

Scaling is performed for several reasons:

1. Improves Model Performance: Many machine learning algorithms perform better when features are on a similar scale.
2. Speeds Up Convergence: Gradient-based algorithms like gradient descent converge faster with scaled data.
3. Prevents Dominance: Prevents features with larger ranges from dominating the learning process.

Difference Between Normalized Scaling and Standardized Scaling

1. Normalized Scaling (Min-Max Scaling):

- Range: Transforms data to a fixed range, usually [0, 1].
- Use Case: Useful when we know the minimum and maximum values of your data and want to scale it to a specific range.

2. Standardized Scaling (Z-score Scaling):

- Range: Transforms data to have a mean of 0 and a standard deviation of 1.
- Use Case: Useful when you want to centre your data and ensure it has unit variance

For instance, in predicting house prices, features like square footage and number of bedrooms have different scales. Without scaling, the larger range of square footage could dominate the model. By applying normalized scaling (rescaling features to a fixed range, usually [0, 1]) or standardized scaling (transforming features to have a mean of 0 and a standard deviation of 1), we ensure that all features contribute equally, improving the accuracy and efficiency of the prediction model.

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

   Answer:

When the Variance Inflation Factor (VIF) is infinite, it typically indicates perfect multicollinearity among the predictor variables in a regression model. This means that one predictor variable is an exact linear combination of one or more other predictor variables.

Here’s why this happens:

1. Perfect Multicollinearity: If one predictor can be perfectly predicted from the others, the denominator in the VIF calculation (which involves the R-squared value of the regression of that predictor on all the other predictors) becomes zero. Since VIF is calculated as 1 / (1 - R^2), a zero denominator results in an infinite VIF.
2. Linear Dependence: In simpler terms, if you have a situation where, for example, X1 = 2*X2 + 3*X3, then X1 is perfectly linearly dependent on X2 and X3. This perfect linear relationship causes the VIF to be infinite.

To address this issue, you can:

1. Remove one of the collinear variables: Identify and remove one of the variables that are causing the perfect multicollinearity.
2. Combine collinear variables: Sometimes, creating a composite variable that combines the collinear variables can help.

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

   Answer:

Q-Q Plot: A graph that compares your data to a normal distribution to see if they match.

Use in Linear Regression

- Check Residuals: In linear regression, we want the errors (residuals) to be normally distributed.
- How to Use: Plot the residuals on the y-axis and the expected normal values on the x-axis.

Importance

- Straight Line: If the points form a straight line, your residuals are normally distributed, which is good.
- Deviations: If the points deviate from the line, it suggests problems like skewness or outliers.

Why It Matters

- Model Accuracy: Ensures your linear regression model is reliable and accurate.

Example Interpretation

- Straight Line: Residuals are normally distributed.
- S-shaped Curve: Indicates skewness.
- Upward/Downward Curves at Ends: Indicates heavy tails

In short, a Q-Q plot helps us to check if our regression model's errors are normally distributed, which is important for the model's validity.