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?

Answer: it seems except weekday all other categorical predictors have significance on bike counts:

- **Season**: Spring has lowest average bikers count and lowest 25th and 75th percentile, whereas fall has highest average bikers count which however is close to average bikers count in summer and winter. The winter average bikers count is a little lower than bikers count in summer.

- **Yr**: Year 2019 has more average bikers count. Even the 75th percentile of bikers count in the year 2018 is just a little higher that 25th percentile of bikers count in the year 2019.

- **mnth**: The average bikers count is lowest in the month of January and highest during the months of June, July and August. The average bikers count increases every month from January until July, thereafter which it starts decreasing till December.

- **holiday**: Holidays have lower average bikers count than oth days. However the 75th percentile is little higher on holidays which could be because of on some holidays people go out for shopping and meet friends and family in larger number than on regular days.

- **weekday**: The weekdays have almost same average counts of bikers across all days with 25th percentile is lowest on Sunday. Which looks obivious as on Sunday people mostly remain at home.

- **workingday**: Working days have higher bikers count than non-working days. However the 75th percentile of bikers on working and non-working days are almost the same.

- **weathersit**: The “Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog” weather has lowest overall bikers count and “Clear, Few clouds, Partly cloudy, Partly cloudy” weather has highest overall bikers count. This is obivious as in “Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog” , no one would want to go out on bikes unless in urgent and unavoidable requirements while “ Clear, Few clouds, Partly cloudy, Partly cloudy” is good for get a bike ride. Meanwhile bikers in “Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist” are 3rd highest and “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds” have 3rd highest bikers count.

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

Answer: Values of a categorical variables are turned up as separate columns. These separate columns are dummy variables. Dummy categorical variables hold boolean values. As the rest of the combinations of values of dummy variables are sufficient to get the dropped “categorical value”, therefore it is recommended to use **drop_first=True** to not create dummy column for first value in a categorical column in order to remove redandacy in the data.

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

with the target variable?

Answer: The 'atemp' and 'temp' look good predictor for 'cnt' as they have higher correlation with 'cnt' than other variables.

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

Answer: After creating the model we get the residual errors and plot the histogram to see error terms, and see if these erros are normally distributed with sum of all values of the errors being zero. We do this with the ‘train’ prediction as well as with the ‘test’ prediction.

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

Answer: By looking at the coefficients of the features in the model, the top 3 contributers having the significance on the prediction of count of bikers are:

1. atemp - it has highest positive correlation with target variable
2. yr - The yr value - 2019 has second highest positive correlation with the target variable
3. weathersit - The value 'Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds' has highest negative correlation with the target variable

General Subjective Questions

Explain the linear regression algorithm in detail.

Answer:

Linear regression is a fundamental supervised learning algorithm used for predictive modeling and regression tasks. It models the relationship between a dependent variable (also called the target or output variable) and one or more independent variables (also called predictors or input features). The algorithm aims to find the best-fitted straight line (in the case of simple linear regression) or hyperplane (in the case of multiple linear regression) that minimizes the errors between the predicted values and the actual target values.

Following sections provide detailed explanation of the linear regression algorithm:

1. **Assumptions**:

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: The observations are independent of each other.
- Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
- Normality: The errors follow a normal distribution with mean 0.

2. **Simple Linear Regression:**

Simple linear regression deals with one dependent variable (Y) and one independent variable (X). The linear regression model is represented as:

$Y = \beta 0 + \beta 1 * X + \varepsilon$

Wherein:

- Y: The dependent variable (target) that we want to predict.
- X: The independent variable (feature) that influences the dependent variable.
- β0 and β1: The intercept and slope of the regression line, respectively.
- ε: The error term representing the difference between the predicted and actual values.

3. **Multiple Linear Regression:**

Multiple linear regression extends the concept to more than one independent variable. The model equation becomes:

$Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + ... + \beta n * Xn + \varepsilon$

Wherein:

- X1, X2, ..., Xn: Multiple independent variables/features.
- β1, β2, ..., βn: Coefficients representing the impact of each independent variable on the dependent variable.

4. **Cost Function:**

The goal of linear regression is to minimize the sum of squared errors (SSE) between the predicted values and the actual values. The cost function (also known as the loss function) is usually the Mean Squared Error (MSE):

$MSE = (1 / N) * \Sigma (Yi – \hat{Y}i)^2$

Wherein:

- N: Number of data points.

- Yi: Actual target value.
- Ŷi: Predicted target value.

5. **Gradient Descent (Optional):**

For large datasets, an optimization technique like gradient descent is used to find the optimal coefficients (β0, β1, ..., βn). Gradient descent iteratively updates the coefficients in the direction that reduces the cost function until it reaches the minimum. This process helps to find the best-fitted line with minimal errors.

6. **Model Evaluation:**

Once the model is trained, it is essential to evaluate its performance. Common evaluation metrics for linear regression models include R-squared, p value, mean absolute error (MAE), and root mean squared error (RMSE).

Linear regression is widely used due to its simplicity, interpretability, and effectiveness in many applications. However, it's important to note that it assumes a linear relationship between the variables and can be sensitive to outliers and violations of its assumptions. For more complex relationships, other regression models like polynomial regression or advanced techniques like regularization (e.g., Lasso, Ridge regression) can be employed.

Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four small datasets that have nearly identical statistical properties but display remarkably different patterns when plotted graphically. These datasets were introduced by the British statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and to caution against relying solely on summary statistics.

The four datasets in Anscombe's quartet share the following statistical properties:

- Number of data points: 11
- Mean of the x-values: approximately 9
- Mean of the y-values: approximately 7.5
- Variance of the x-values: approximately 11
- Variance of the y-values: approximately 4.125
- Correlation between x and y: approximately 0.816

Now, let's explore each dataset in Anscombe's quartet:

Dataset 1:

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84

7	4.82
5	5.68

Dataset 2:

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

Dataset 3:

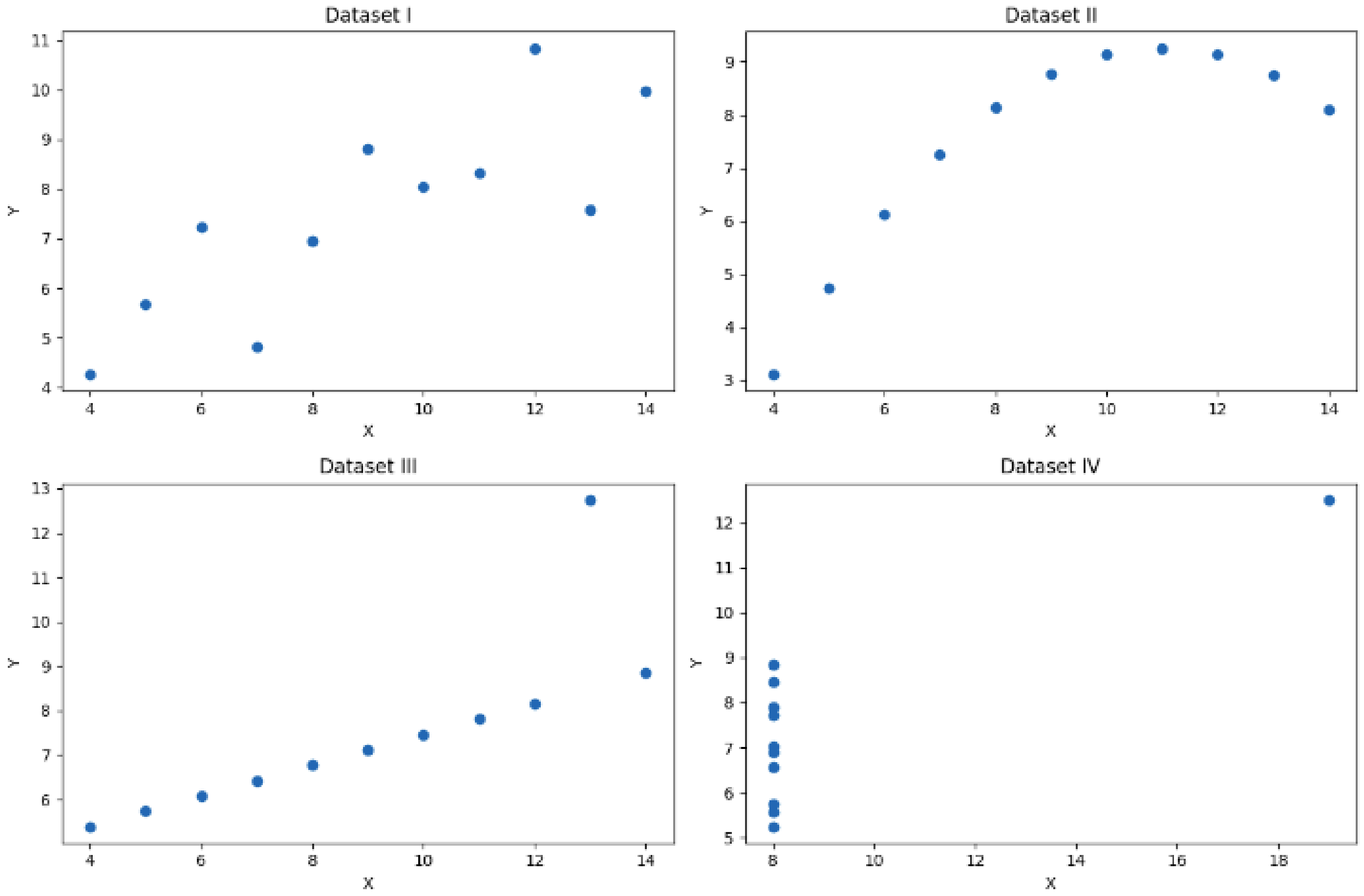
X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Dataset 4:

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04

8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

While all four datasets share similar summary statistics, their graphical representations differ significantly. When plotted as scatter plots or line plots, each dataset exhibits different patterns, ranging from linear relationships to curvilinear relationships and even cases with no apparent relationship between x and y:



The primary lesson from Anscombe's quartet is that solely relying on numerical summaries like means, variances, and correlations can be misleading. Visualizing data is essential to gain a deeper understanding of the underlying patterns and relationships within the data. Effective data visualization can help identify outliers, non-linear relationships, and potential issues with the data, providing valuable insights that might not be apparent from summary statistics alone.

What is Pearson’s R?

Pearson's correlation coefficient, often denoted as Pearson's R or simply as "r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson, a British mathematician and biostatistician, in the late 19th century. Pearson's R is one of the most widely used correlation coefficients in statistics and is a key tool for understanding the association between two variables.

The Pearson correlation coefficient ranges from -1 to +1:

- If r = +1, it indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.

- If r = -1, it indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.

- If r = 0, it indicates no linear relationship, suggesting that the variables are not correlated.

The formula to calculate Pearson's correlation coefficient (r) between two variables X and Y, each with n data points, is as follows:

$r = (\Sigma((Xi - \bar{X}) * (Yi - \bar{Y}))) / (n * \sigma X * \sigma Y)$

Where:

- Xi, Yi: Individual data points for X and Y, respectively.

- $\bar{X}$, $\bar{Y}$: Mean of X and Y, respectively.

- σX, σY: Standard deviation of X and Y, respectively.

The calculation involves finding the covariance between X and Y ($\Sigma((Xi - \bar{X}) * (Yi - \bar{Y}))$) and normalizing it by the product of the standard deviations of X and Y.

Interpreting Pearson's R:

- If r is close to +1, it suggests a strong positive linear relationship.

- If r is close to -1, it indicates a strong negative linear relationship.

- If r is close to 0, it implies weak or no linear relationship.

It's important to note that Pearson's correlation coefficient measures only linear relationships between variables. If the relationship is non-linear, Pearson's R may not accurately capture the association. In such cases, other correlation measures, like Spearman's rank correlation or Kendall's tau, may be more appropriate.

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

Scaling is a preprocessing step in data preparation that involves transforming numerical features to a specific range or distribution. The goal of scaling is to bring all features to a common scale, which can improve the performance and convergence of various machine learning algorithms.

Why Scaling is Performed:

1. **Avoiding Bias**: Many machine learning algorithms, such as gradient-based optimization methods, are sensitive to the scale of features. Features with larger scales can dominate the learning process and lead to biased model predictions.

2. **Accelerating Convergence**: Scaling can help gradient-based optimization converge faster, as the algorithm can take more balanced steps toward the optimal solution when features are on a similar scale.

3. **Improving Model Performance**: Scaling can lead to improved model performance, especially for distance-based algorithms like k-nearest neighbors (KNN) or support vector machines (SVM), which rely on the distances between data points.

Normalized Scaling vs. Standardized Scaling:

Normalized Scaling (also known as Min-Max scaling) and **Standardized Scaling** (also known as Z-score scaling or Standardization) are two common scaling techniques:

1. **Normalized Scaling (Min-Max scaling):**

- Normalized scaling scales the data to a fixed range, typically [0, 1]. It preserves the shape of the original distribution but shifts and scales it to fit within the specified range.
- The formula for normalized scaling is:

X_normalized = (X - X_min) / (X_max - X_min)

- X: Original feature values
- X_min: Minimum value of the feature
- X_max: Maximum value of the feature

2. **Standardized Scaling (Z-score scaling):**

- Standardized scaling transforms the data to have zero mean and unit variance. It centers the data around the mean and scales it by the standard deviation.
- The formula for standardized scaling is:

X_standardized = (X - X_mean) / X_std

- X: Original feature values
- X_mean: Mean of the feature
- X_std: Standard deviation of the feature

Comparison:

- Normalized Scaling: Useful when you have a defined range in which you want your data to be mapped (e.g., [0, 1]). It does not change the distribution's shape and is suitable for algorithms that expect data in a bounded range.

- Standardized Scaling: Useful when you want your data to have zero mean and unit variance. It standardizes the data and is helpful for algorithms that rely on the relative distances between data points.

In summary, choose the appropriate scaling method based on the requirements and characteristics of your data and the machine learning algorithm you intend to use.

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

The formula to calculate the VIF for each variable i is:

VIF_i = 1 / (1 – R^2_i)

Clearly if R squared is equal to 1 then the VIF would go to infinite. This means following:

The situation of perfect multicollinearity, where one variable is an exact linear combination of others, can lead to numerical instability in the model's calculations, and it becomes impossible to estimate unique coefficients for all variables involved.

To address this issue, one of the redundant variables can be removed from the model. This process is called "feature selection" or "dimensionality reduction."

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess the distributional similarity between a given dataset and a theoretical distribution, typically the normal distribution. It is a useful visual technique to check if a dataset follows a specific distribution or to identify departures from the assumed distribution.

The Q-Q plot compares the quantiles of the dataset against the quantiles of the theoretical distribution. If the dataset follows the theoretical distribution, the points in the Q-Q plot will approximately lie on a straight line. The x-axis represents the theoretical quantiles, and the y-axis represents the sample quantiles. The closer the points are to the straight line, the better the data aligns with the theoretical distribution.

Use and Importance of Q-Q Plot in Linear Regression:

In linear regression, the Q-Q plot is an essential diagnostic tool to assess the assumption of normality of the residuals. The residuals are the differences between the observed dependent variable values and the predicted values from the linear regression model.

The normality assumption is crucial for the following reasons:

Statistical Inference: Many statistical tests, confidence intervals, and p-values used in linear regression assume that the residuals are normally distributed. Violating this assumption may lead to incorrect inferences.

Model Accuracy: Normally distributed residuals indicate that the model's errors have consistent variance across all predicted values, which is a key assumption for linear regression.

Model Interpretation: When residuals are normally distributed, the coefficients' estimates become the best linear unbiased estimates.

By plotting the residuals on a Q-Q plot, you can visually inspect whether they follow a normal distribution. If the points in the Q-Q plot deviate significantly from the straight line, it suggests that the residuals are not normally distributed, and the normality assumption may be violated.

Additionally, the Q-Q plot is not limited to assessing normality assumptions; it can be used to check the distributional fit to other theoretical distributions as well, such as the exponential or log-normal distribution.

In summary, the Q-Q plot is a required tool in linear regression to check the normality assumption of the residuals, which, in turn, affects the validity and accuracy of the regression model's estimates and inferences.