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 –

We have 2 categorical variables “season” and “weathersit”. After analysis we can say that

1) Throughout all the seasons’ count of total bikes rented remains somewhat constan.
2) While for weathersit count of total bikes rented is higher when weather is clear and low when it’s cloudy or have light rain that particular day.

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

Answer –

We use dummy variables to accommodate categorical variables. While dummy_variable function will create **n** variables to represent a categorical variable with **n** unique values, though it can be represented perfectly by **n-1** variables.

This ensures optimal dataset variables and avoid unnecessary variables to deal with.

By default dummy_variable function will create n variable columns for a categorical variable. Using **drop_first = TRUE** ensures that total variables created are n-1.

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

Answer –

temp variable has highest correlation with target variable **cnt** among the numerical variables.

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

Answer –

1) By plotting distribution plot for residual term (y_train-y_pred) and verifying it has normal distribution. 2) By calculating VIF (Variance Inflation Factor) values for multicollinearity. If VIF>5 , Extreme Multicollinearity 3) By checking Durbin-Watson value to check autocorrelation.

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

Answer – **temp**, **light_rain** and **yr**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer –

1) Linear regression is a statistical method that is used to predict a continuous dependent variable(target variable) based on one or more independent variables(predictor variables). This technique assumes a linear relationship between the dependent and independent variables, which implies that the dependent variable changes proportionally with changes in the independent variables. In other words, linear regression is used to determine the extent to which one or more variables can predict the value of the dependent variable.

2) Linear regression is a supervised algorithm that learns to model a dependent variable, y as a function of some independent variables (aka "features") X, by finding a line (or surface) that best "fits" the data. In general, we assume y to be some number and each Xi can be basically anything. For example: predicting the price of a house using the number of rooms in that house (y: price, X1: number of rooms) or predicting weight from height and age (y: weight, X1: height, X2: age).

In general, the equation for linear regression is

$y=\beta0+\beta1*X1+\beta2*X2+...+\beta p*Xp+c$

where:

- y: the dependent variable; the thing we are trying to predict.
- Xi: the independent variables: the features our model uses to model y.
- βi: the coefficients (aka "weights") of our regression model. These are the foundations of our model. They are what our model "learns" during optimization.
- c: the constant/irreducible error in our model. A term that collects together all the unmodeled parts of our data.

3) Assumptions We Make in a Linear Regression Model:
   - The relationship between dependent and independent variables should be linear.
   - No multicollinearity in the data. Multicollinearity occurs when independent variables are not independent of each other.
   - Error terms are normally distributed
   - Error terms have constant variance (homoscedasticity)
4) There are two main types of linear regression:
   - Simple linear regression
     Simple linear regression is an approach for predicting a response using a single feature.
   - Multiple linear regression
     Multiple linear regression attempts to model the relationship between two or more features

2. Explain the Anscombe's quartet in detail.

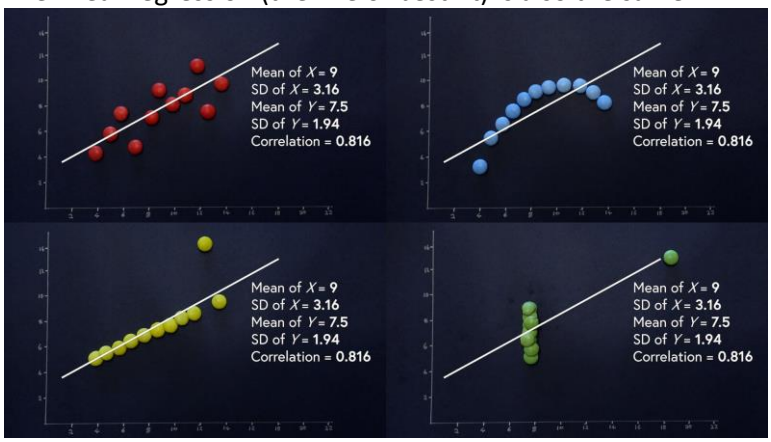
Answer –

Summary metrics can be misleading, because they reduce complex patterns in your data down to simple, singular figures. As a result, you can miss important trends that are only ever revealed when you graph your data in the form of a data visualization.
Anscombe's Quartet shows how four entirely different data sets can be reduced down to the same summary metrics.
Anscombes quartet is a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Dataset

Red		Blue		Yellow		Green	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

When summary metrics are calculated, datasets look practically identical:
Mean of X = 9, **Standard deviation of X** = 3.16, **Mean of Y** = 7.5, **Standard deviation of Y** = 1.94 and **Correlation between X & Y**= 0.816
The linear regression (the line of best fit) is also the same:



But by looking at Scatter plot you can tell that all 4 datasets are different.
It means that if you don't want to miss important trends, you need to visualize your data and break it down.
In **conclusion** summary metrics can provide a helpful snapshot of business performance, but they can also mislead. The best way to see what's going on with your business KPIs is to visualize them, so you have a clearer picture of how and why your metrics are changing.

3. What is Pearson's R?

Answer –
Pearson's correlation coefficient (r) is a measure of the linear association of two variables.

Correlation analysis usually starts with a graphical representation of the relation of data pairs using a scatter plot. Pearsons r a measure that ranges from -1 to +1, where:
1) +1 indicates a perfect positive correlation,
2) -1 indicates a perfect negative correlation, and
3) 0 indicates no correlation at all.
The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when 1) Both variables are quantitative, 2) The variables are normally distributed, 3) The data have no outliers and 4) The relationship is linear.
Pearson's r is also know as Bivariate correlation, Pearson product-moment correlation coefficient (PPMCC) and The correlation coefficient

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

Answer –
1) Sometime the raw dataset has variable values which are spread over large range and can affect the model adversely. Scaling is method to adjust the spread or variability of the data.
2) Robustness to Outliers: Scaling can make your models less sensitive to extreme values and Algorithm Compatibility: Some algorithms, like Support Vector Machines and Principal Component Analysis, work best with scaled data.
3) **Normalization Scaling**
Normalization or Min-Max Scaling is used to transform features to be on a similar scale. This scales the range to [0, 1] or sometimes [-1, 1]. Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale features like age and not incomes because only a few people have high incomes but the age is close to uniform.
Standardization scaling
Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score. Standardization Scales features to have a mean of 0 and a standard deviation of 1. Standardization can be helpful in cases where the data follows a normal distribution. However, this does not have to be necessarily true. Standardization does not get affected by outliers because there is no predefined range of transformed features.

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

Answer –
Infinite VIF values in Python typically occur due to perfect multicollinearity in your data.
Reasons:

- Perfect Correlation: If two or more independent variables are perfectly correlated (correlation coefficient of 1 or -1), it becomes impossible to estimate their individual effects on the dependent variable. This leads to infinite VIF values.
- Redundant Variables: If one independent variable is a linear combination of other independent variables, it adds no new information to the model. This is another form of perfect multicollinearity and results in infinite VIFs.
- Including a constant term multiple times: If you accidentally include the same constant variable multiple times in your model, it creates perfect multicollinearity and infinite VIFs.