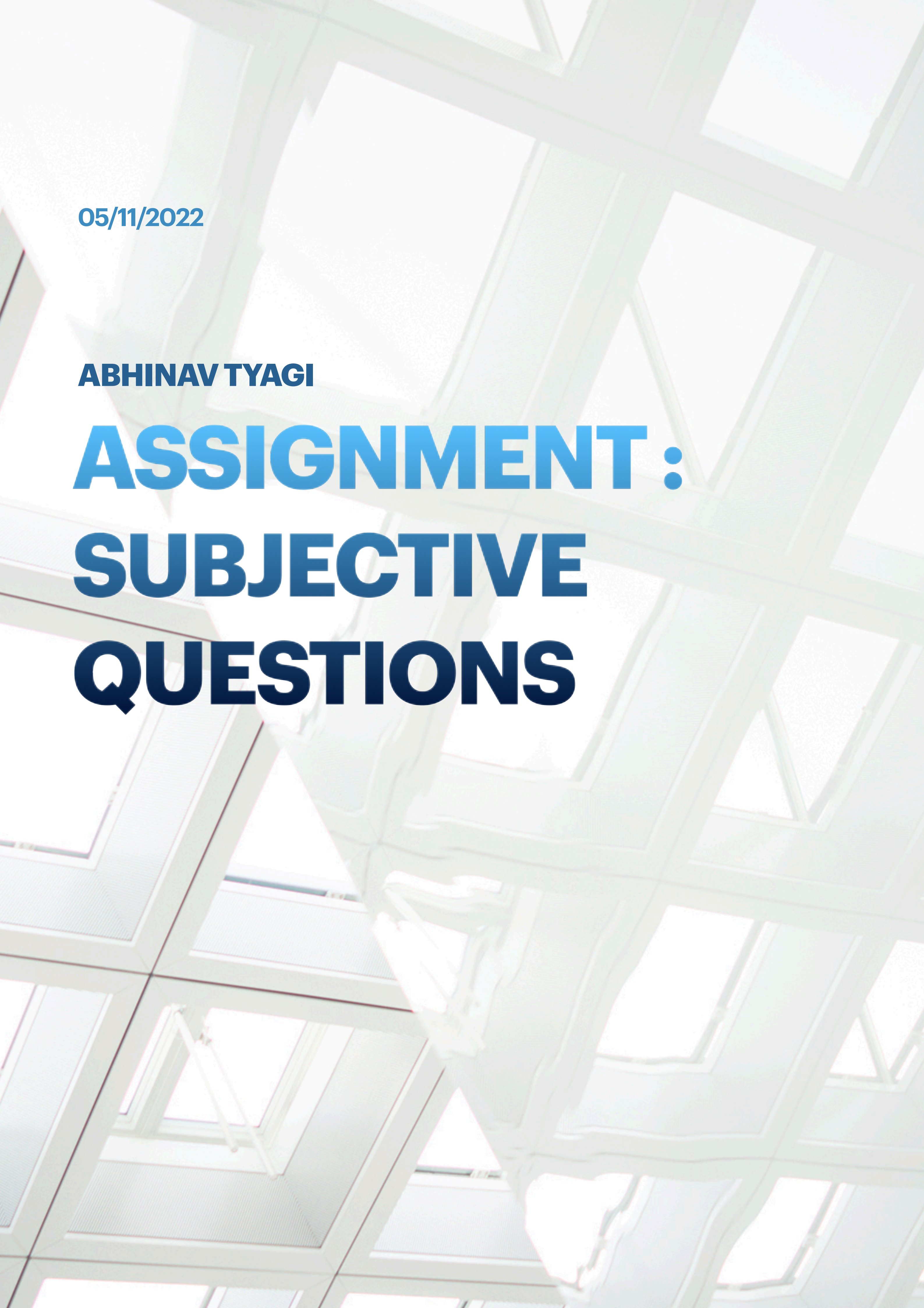


05/11/2022

ABHINAV TYAGI

ASSIGNMENT : SUBJECTIVE QUESTIONS

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? (3 marks)**

Final Report

Train R-Squared	*0.824*
Train R-Squared Adjusted	*0.821*
Test R-Squared	*0.820*
Test R-Squared Adjusted	*0.812*

Top 3 Predictor Variables

Variable	Relation
temp (Temperature)	*Per Unit Increase in temp yields bike bookings raised by 0.563615 times.*
weathersit_3 (Weather Situation 3)	*Per Unit Increase in weathersit_3 yields bike bookings decreased by -0.306992 times.*
yr (Year)	*Per Unit Increase in yr yields bike bookings raised by 0.230846 times.*

Inferences:

A. Temperature increases (correlated with summers, a pleasant time in the US), leads more people outdoor for adventure. Therefore more rentals.
B. Weather Situation 3, Meaning light snow and showers, makes more people avoid 2 wheeler travelling, therefore a strong negative correlation with rise in rains and light snow.
C. Logical, as each year business grows.

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

We do NOT need another column for "Uknown".

It can be necessary for some situations, while not applicable for others. The goal is to reduce the number of columns by dropping the column that is not necessary. However, it is not always true. For some situations, we need to keep the first column.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Temp has the highest correlation with cnt.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

General Subjective Questions

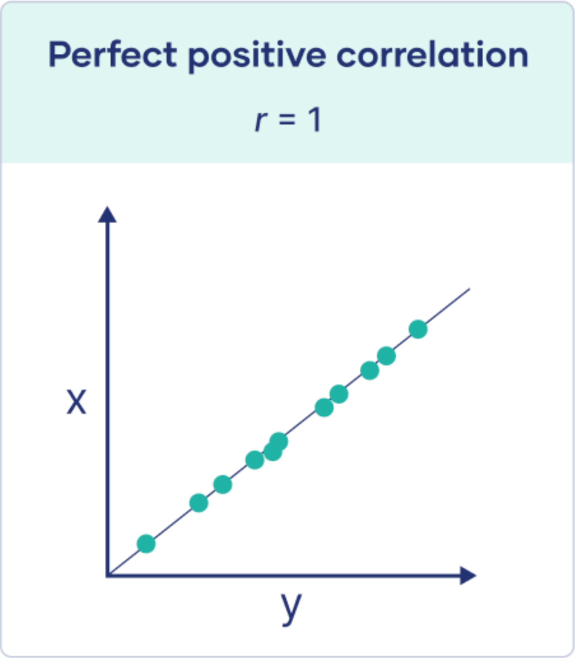
1. **Explain the linear regression algorithm in detail. (4 marks)**

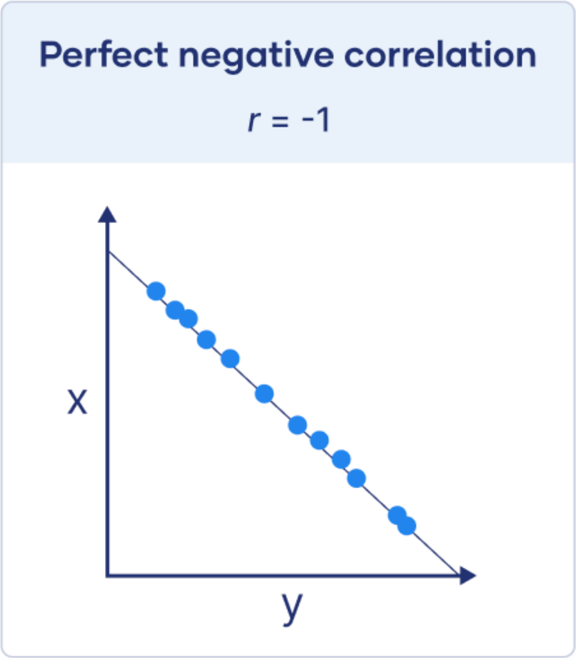
2. **Explain the Anscombe's quartet in detail. (3 marks)**

3. **What is Pearson's R? (3 marks)**

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.





4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**