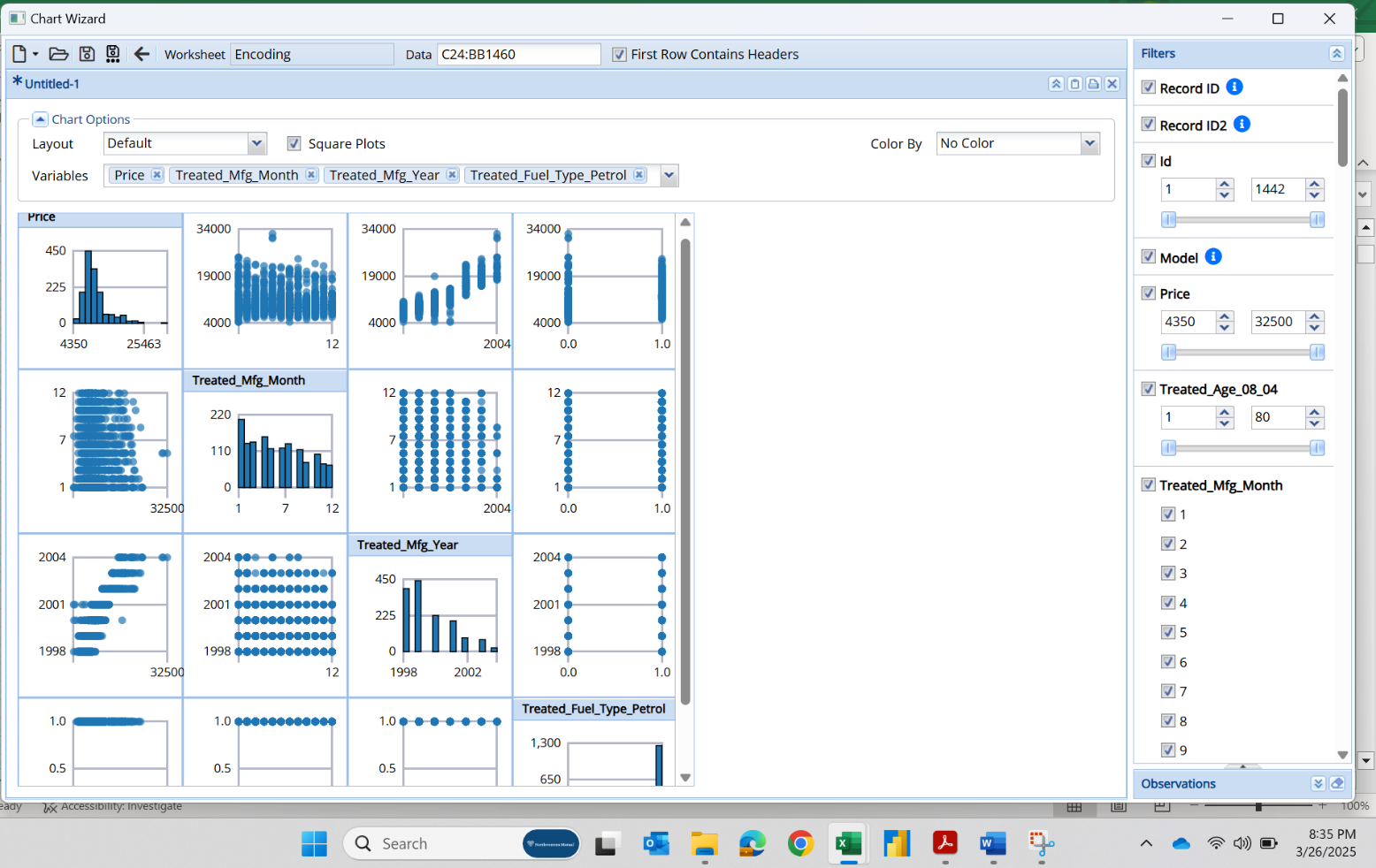
**Assignment 2 – Principal Component Analysis**

1. How did you clean, explore the data? Were there any categorical variables in the dataset? Explain the scatter plot matrix results in terms of correlation amongst predictors and with the outcome. Provide screen shot of the matrix.

The data was first cleaned by employing missing data handling, followed by transforming categorical text columns into dummy variables. The two categorical variables that were transformed are the fuel type and color column.



For the predictors, price and manufacturing year have a strong positive correlation, which is expected. Notice the manufacturing month and manufacturing year have no clear correlation. The correlation between price and having a petroleum fuel type has no correlation, which follows because most cars ran on petroleum in 2004.

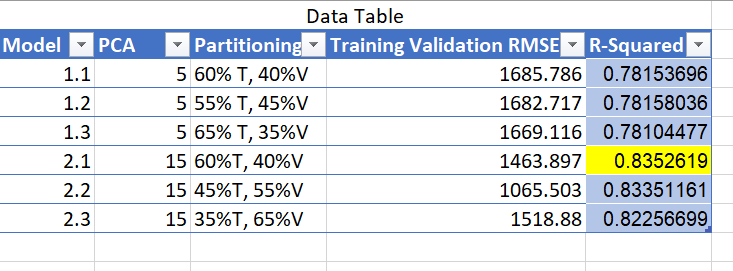
* 1. 2. Which top 5 variables did you select using PCA technique and which Smallest # gave you max. variance, how much variance did the Top 15 actually cover?

For the PCA technique, the top 5 variables selected were Treated\_Age\_08\_04, Treated\_Fuel\_Type\_Petrol, Treated\_Backseat\_Divider, Treated\_Mistlamps, Treated\_Radio.

**The Top 15 covered approximately 64.49% of the variance.**

3. How did you select the best model - explain the ‘full selection criteria’ you followed for comparison between the two models? (Refer to the Lecture 2 slides for ‘full model selection criteria’ for Linear Regression)

**Prepare a Table showing** – all the models, PCA, Partitioning, Training Validation RMSE, R-squared



The best model was the 15 principal component variable model with the 60/40 Training to Validation ratio. The deciding factor in model selection is the r-squared value, the highest value is desired. In this case it would be **model 2.1**.

[Toyota Sales - Model 1.1.xlsx](https://nmfco-my.sharepoint.com/:x:/g/personal/davaskar_barron_nm_com/EbktVlrJR-RLn0bI0aYi6dwBCz1_gPuiDt5rGyYMJsrzSQ?e=Evrfty)

[Toyota Sales - Model 1.2.xlsx](https://nmfco-my.sharepoint.com/:x:/g/personal/davaskar_barron_nm_com/EXP_9R75m_JAnU7YcvCtKV0BlqPOC0SoiK8QOhr4wi8SiA?e=g4yafv)

[Toyota Sales - Model 1.3.xlsx](https://nmfco-my.sharepoint.com/:x:/g/personal/davaskar_barron_nm_com/EUQbe0s0ctFNtEkGN7pNU3gB0OlERVhDeX0YqSMVSd52tA?e=rPZqjk)

[Toyota Sales - Model 2.1.xlsx](https://nmfco-my.sharepoint.com/:x:/g/personal/davaskar_barron_nm_com/EbyXZG6KoeFLtpFSVkRwFEQBBfoNiK26neGdnEP2Y3D0og?e=7zsaRO)

[Toyota Sales - Model 2.2.xlsx](https://nmfco-my.sharepoint.com/:x:/g/personal/davaskar_barron_nm_com/EYNAO4jBkIVNmKgMDggD5m0BmzO1ElnIPAvNv2XIiSwYsA?e=oG9ekN)

[Toyota Sales - Model 2.3.xlsx](https://nmfco-my.sharepoint.com/:x:/g/personal/davaskar_barron_nm_com/EdqLhLy__X1GsltLgW2t9xQBOuO1Z0tTYmJ9Ts_Sj7jSYA?e=6KpySR)