Statistical Assessment: Automobile Price Analysis Study

Chiranjeevi Nalapalu, PhD

Executive Summary:

1. A strong correlation exists between vehicle horsepower and engine size with car price (correlation coefficients > 0.8). Ordinary Least Squares (OLS) regression analysis demonstrates that engine size significantly impacts price (p-value < 0.05).

2. Significant price differences were observed across vehicle body types (p-value < 0.05), whereas no statistically significant price variations were detected among different fuel types (p-value > 0.05).

3. No significant relationship was found between car body types and their respective price categories (Chi-Square test p-value > 0.05).

4. The multiple linear regression model developed to predict car prices yielded strong predictive power, with an R^2 value of approximately 0.83 when evaluated on the test dataset after train-test splitting.

5. All numerical features were evaluated for predictive capability. Engine size, horsepower, curb weight, and car width emerged as the most influential variables in the regression analysis as they had the highest coorelation to Price.

6. Sales were forecasted for Q1 2025 using two time series models with a projected value of 577 units

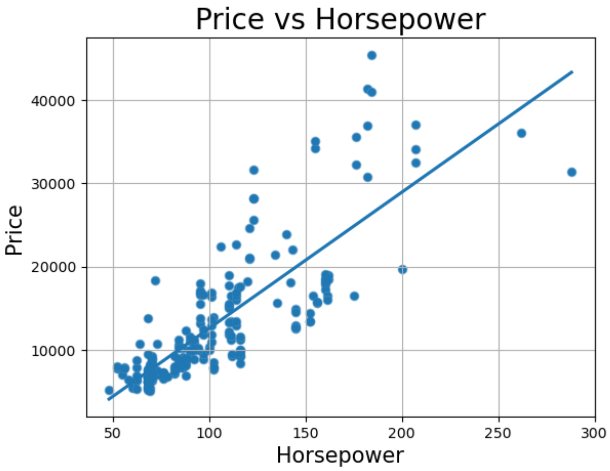
7. A weak relationship was identified between both price and advertising expenditure on sales, though these relationships were not statistically significant (p-value > 0.05).

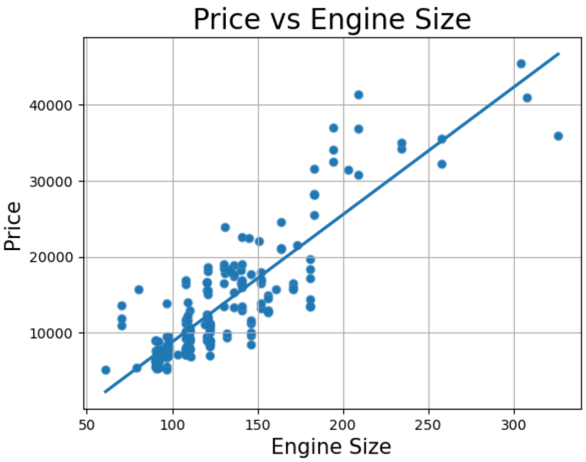
Question 1:

Describe how you would evaluate the correlation between horsepower, engine size, and price and find whether engine size significantly impacts price. What statistical test would you use, and why?

To evaluate the correlation between horsepower, engine size, and price, both Pearson and Spearman correlation coefficients were used. Both features are strongly and positively coorelated to Car Price as evident from the results of statistical tests and plots below.

	Pearson Correllation Coefficient (p-value)	Spearman Correllation Coefficient (p-value)
Horsepower vs Price	0.81 ($1.4x10^{-48}$)	0.85 ($1.4x10^{-59}$)
Engine Size vs Price)	0.87 ($1.4x10^{-65}$)	0.83 ($1.4x10^{-52}$)





A OLS Regression analysis was conducted to determine if the engine size significatly imapcts price. Model summary is as follows

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.764
Model:                            OLS   Adj. R-squared:                  0.763
Method:                 Least Squares   F-statistic:                     657.6
Date:                Tue, 29 Apr 2025   Prob (F-statistic):           1.35e-65
Time:                        16:54:03   Log-Likelihood:                -1984.4
No. Observations:                 205   AIC:                             3973.
Df Residuals:                     203   BIC:                             3979.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -8005.4455    873.221     -9.168      0.000   -9727.191   -6283.700
enginesize   167.6984      6.539     25.645      0.000     154.805     180.592
==============================================================================
Omnibus:                       23.788   Durbin-Watson:                   0.768
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               33.092
Skew:                           0.717   Prob(JB):                     6.52e-08
Kurtosis:                       4.348   Cond. No.                         429.
==============================================================================
```

The analysis revealed a statistically significant impact of engine size on price. This significance is supported by a p-value less than 0.05.

Furthermore, the R-squared value is 0.764, indicating that approximately 76.4% of the variance in car prices can be explained by engine size.

The F-statistic of 657.6, with a corresponding p-value of 1.35e-65, confirms that the overall model is statistically significant.

In summary, these results indicate a strong positive relationship between engine size and car price, where engine size is a significant predictor of car price.

Question 2:

Propose a method to analyse if there is a difference in price across different car body types and price across fuel types. And interpret your findings

To analyze price differences across different car body types and fuel types, an ANOVA test was employed. The data was fitted to an Ordinary Least Squares (OLS) regression model, and an ANOVA table was generated. The key results from the ANOVA table are summarized below:

Car body type:

```
                  sum_sq     df         F    PR(>F)
C(carbody)  1.801997e+09    4.0  8.031976  0.000005
Residual    1.121764e+10  200.0       NaN       NaN
```

Fuel Type:

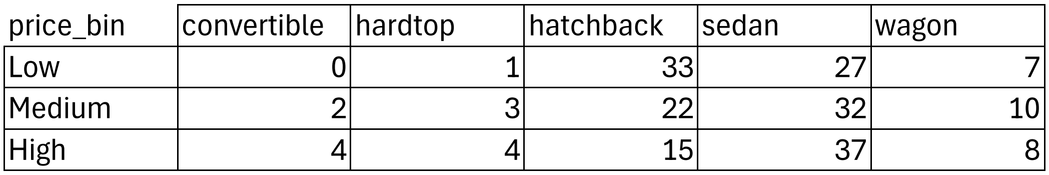
```
                   sum_sq     df         F    PR(>F)
C(fueltype)  1.454053e+08    1.0  2.292741  0.131536
Residual     1.287423e+10  203.0       NaN       NaN
```

P-values indicate significant price variations between different Car body types (p-value < 0.05), But the same is not the case with different fuel types which has no significant relation to price(p-value > 0.05).

<u>Question 3:</u>

Find out the relationship between price and car body types. (Use Chi-Square)

To determine the relationship between car prices and body types, the prices were categorized into three quantiles: "Low," "Medium," and "High." A contingency table was constructed to compare price categories with car body types, as shown below:



price_bin	convertible	hardtop	hatchback	sedan	wagon
Low	0	1	33	27	7
Medium	2	3	22	32	10
High	4	4	15	37	8

Chi-Square test was conducted using this data, yielding the following results:

- **Chi-Square Value**: 15.00
- **Degrees of Freedom (dof)**: 8
- **p-value**: 0.06

Based on the Chi-Square test, no significant relationship was identified between car body types and price categories, as the p-value is greater than 0.05.

Question 4:

What approach would you take to build a model predicting price based on the dataset?Outline the steps and explain the challenges you have faced

Steps Taken:

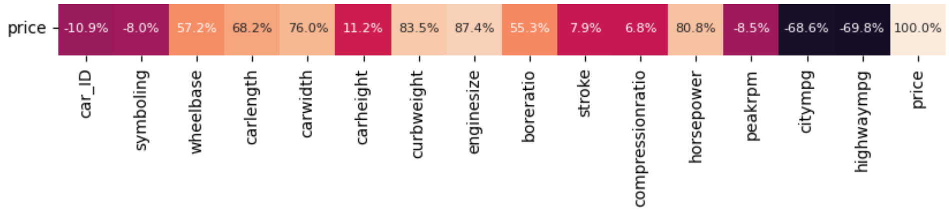
1. Observing the linear relationship between horsepower, engine size, and price, as well as analyzing the correlation heatmap, I concluded that a Linear Regression (LR) model would be most suitable for predicting price.
2. Non-numerical data was removed, and the dataset was split into training and test sets.
3. Initially, a simple LR model was built using only one feature: engine size (which had the highest correlation to price). This approach yielded promising results with an R^2 score of 0.78 and a Mean Squared Error (MSE) of 3341.
4. Next, a multiple LR model was developed using all numerical features. This enhanced the results, achieving an R^2 score of 0.82 and an MSE of 2595.
5. The multiple LR model was then refined by selecting only the features with the highest correlation to price (correlation coefficient > 0.6). This optimization resulted in a similar performance, with an R^2 score of 0.80 and an MSE of 2914.

Challenges Encountered:

- Deciding whether to deepen domain knowledge to better understand the importance of specific parameters.
- Exploring the potential of feature engineering to further improve prediction accuracy. However, the simple model produced satisfactory results, leading me to halt further iterations.

Question 5:

What features would you consider most important for predicting price, and how would you validate this selection?



Key Features: The features identified as most important for predicting price are:

- Engine Size
- Horsepower
- Curb Weight
- Car Width
- Car Length
- City MPG
- Highway MPG

These features displayed the highest correlation with price, making them strong predictors.

Validation: As mentioned earlier, a Multiple Linear Regression (LR) model was run using these selected features. The model demonstrated improved performance with:

- **R² Score**: 0.80
- **Mean Squared Error (MSE)**: 2914

This validates the effectiveness of these features in accurately predicting car prices.

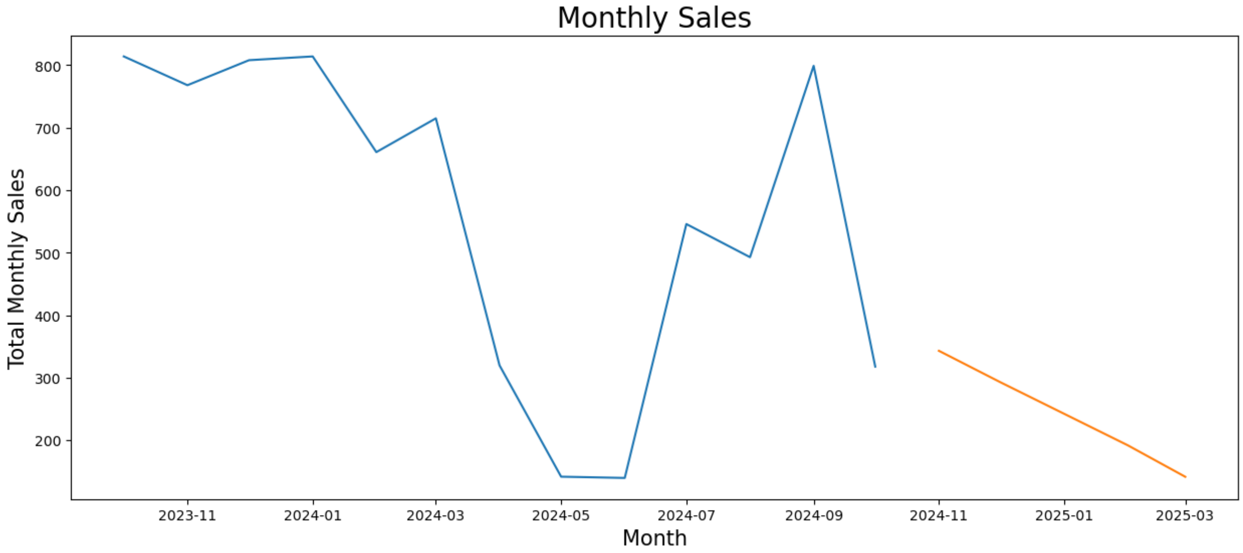
Question 6:

Find car sales data in CarAssignment2, Give the best sales forecast value for the next quarter (From Jan to Mar). You can use multiple models, and justify your forecast value as the best.

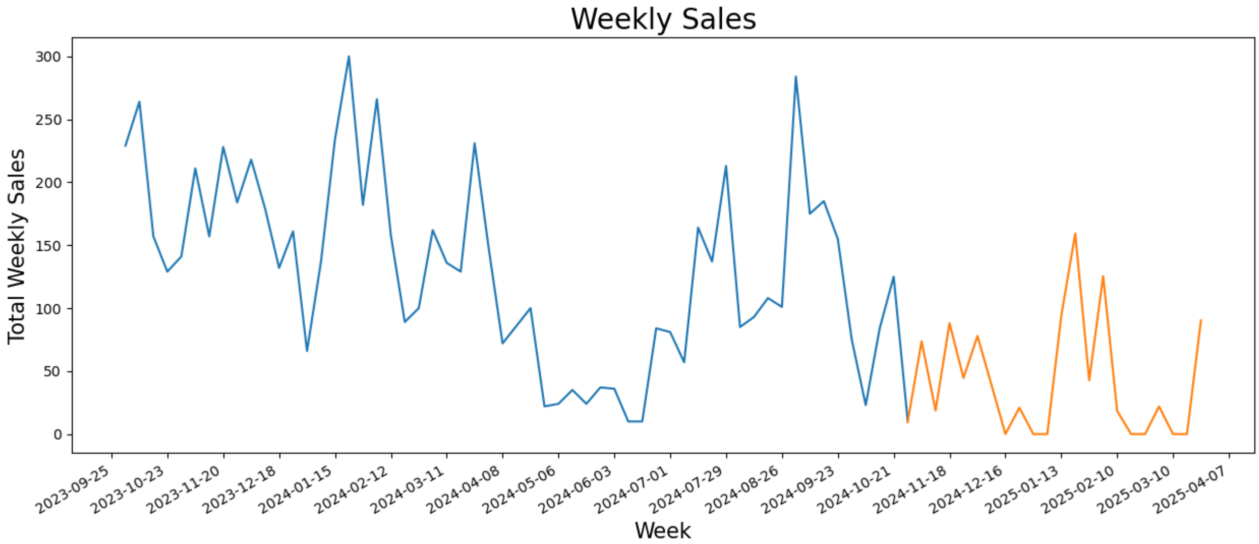
Approach and Analysis: The sales data, originally presented in daily amounts, showed inconsistent patterns, with multiple sales on some days and no sales on others. To address this, the data was grouped into different time intervals: days, weeks, and months. A time series plot revealed that the daily scale was too noisy to build a reliable model. Thus, only weekly and monthly grouped data were used for forecasting.

Models Used:

1. **Monthly Data**:
   - A simple Exponential Smoothing model was applied, given that only 13 data points were available.
   - Forecast for the months of January to March (Q1 2025): **577 units**.



2. **Weekly Data**:
    - A Seasonal ARIMAX model was used, leveraging the 56 data points available.
    - Seasonal adjustments were incorporated, and negative values were removed.
    - Forecast for Q1 2025: **552 units**.



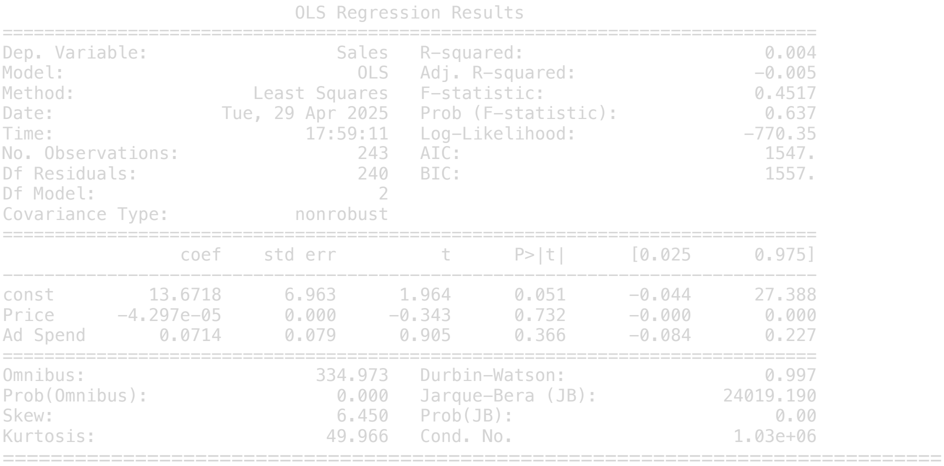
Justification of the Best Forecast Value: Both models provided reasonable forecasts based on their respective datasets. The Exponential Smoothing model relied on a smaller dataset but captured broader trends, while the Seasonal ARIMAX model incorporated seasonality and handled more detailed weekly patterns.

Ultimately, the forecast from the Exponential Smoothing model, **577 units**, may be preferred for its simplicity and reliance on aggregated monthly data, which smooths out short-term fluctuations. However, the choice between the two depends on the specific context and the accuracy requirements of the forecast.

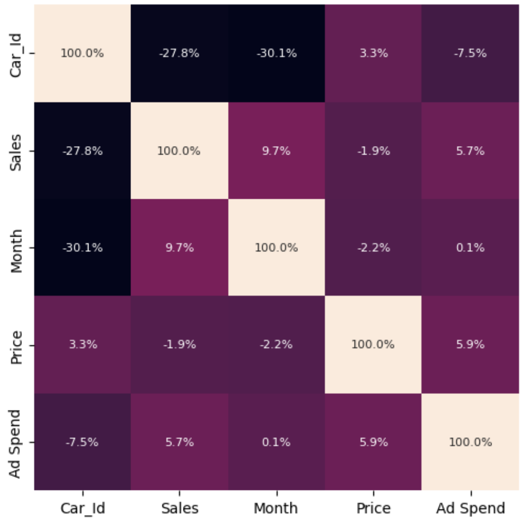
Question 7:

Find the relationship between price and ad spend on sales using data in a sheet named CarAssignment3.

Analysis: An OLS Regression analysis was conducted to examine whether Price and Ad Spend significantly impact Sales. The summary of the results is as follows:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                 -0.005
Method:                 Least Squares   F-statistic:                    0.4517
Date:                Tue, 29 Apr 2025   Prob (F-statistic):              0.637
Time:                        17:59:11   Log-Likelihood:                -770.35
No. Observations:                 243   AIC:                             1547.
Df Residuals:                     240   BIC:                             1557.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.6718      6.963      1.964      0.051      -0.044      27.388
Price      -4.297e-05      0.000     -0.343      0.732      -0.000       0.000
Ad Spend       0.0714      0.079      0.905      0.366      -0.084       0.227
==============================================================================
Omnibus:                      334.973   Durbin-Watson:                   0.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            24019.190
Skew:                           6.450   Prob(JB):                         0.00
Kurtosis:                      49.966   Cond. No.                     1.03e+06
==========================================================================================
```



	Car_Id	Sales	Month	Price	Ad Spend
Car_Id	100.0%	-27.8%	-30.1%	3.3%	-7.5%
Sales	-27.8%	100.0%	9.7%	-1.9%	5.7%
Month	-30.1%	9.7%	100.0%	-2.2%	0.1%
Price	3.3%	-1.9%	-2.2%	100.0%	5.9%
Ad Spend	-7.5%	5.7%	0.1%	5.9%	100.0%

Findings:

- Both **Price** and **Ad Spend** have p-values > 0.05, indicating that neither feature has a significant relationship with Sales.
- The R-squared value (0.004) suggests that the model explains only 0.4% of the variation in Sales, reflecting a very weak relationship.

Additional Insights: A correlation heatmap was also generated, revealing minimal correlation between Price, Ad Spend, and Sales, which aligns with the regression findings. These results confirm that neither Price nor Ad Spend serves as a strong predictor of Car Sales.