COURSEWORK BUSINESS REPORT:

VISUALIZATION AND STATISTICAL ANALYSIS OF A REALWORLD DATA SET

STUDENT NAME - Atul Kumar Meena

STUDENT NUMBER - 240038227

TITLE - Market Analysis of Second-hand cars

MODEL NAME - Toyota C-HR

Email - atuljeph03@gmail.com

Prepared with - PYTHON , SPSS

Aston University
BIRMINGHAM UK

TABLE OF CONTENT

INTRODUCTION

We discovered that part of the data comprised null values, or blank values, which may have an effect on the predictive model's accuracy after obtaining a sizable amount of data roughly 300 population data points from the auto trader resource website. In order to appropriately evaluate and analyze the results, we therefore logically shaped the data.

OUR STRATEGY

To forecast the prices of pre-owned vehicles, information on various aspects such as the brand and model of the car, its age, mileage, and overall state can be accumulated and used to construct a predictive model. This model can subsequently be employed to approximate the worth of a specific used car, considering its attributes and the present market dynamics.

OVERALL ANALYSIS

This analysis assesses and predicts the valuation of a Toyota C-HR B170PY through regression analysis, taking into account factors such as vehicle age, mileage, fuel efficiency, transmission method, and production year. The report highlights the pivotal aspects of engine performance and semi-automatic transmission, intending to offer strategic insights for stakeholders in the pre-owned automobile sector, whether they are prospective purchasers or sellers.

- **The Toyota C-HR that I have chosen is for the postcode B170PY.**

- **Source:** https://www.autotrader.co.uk/

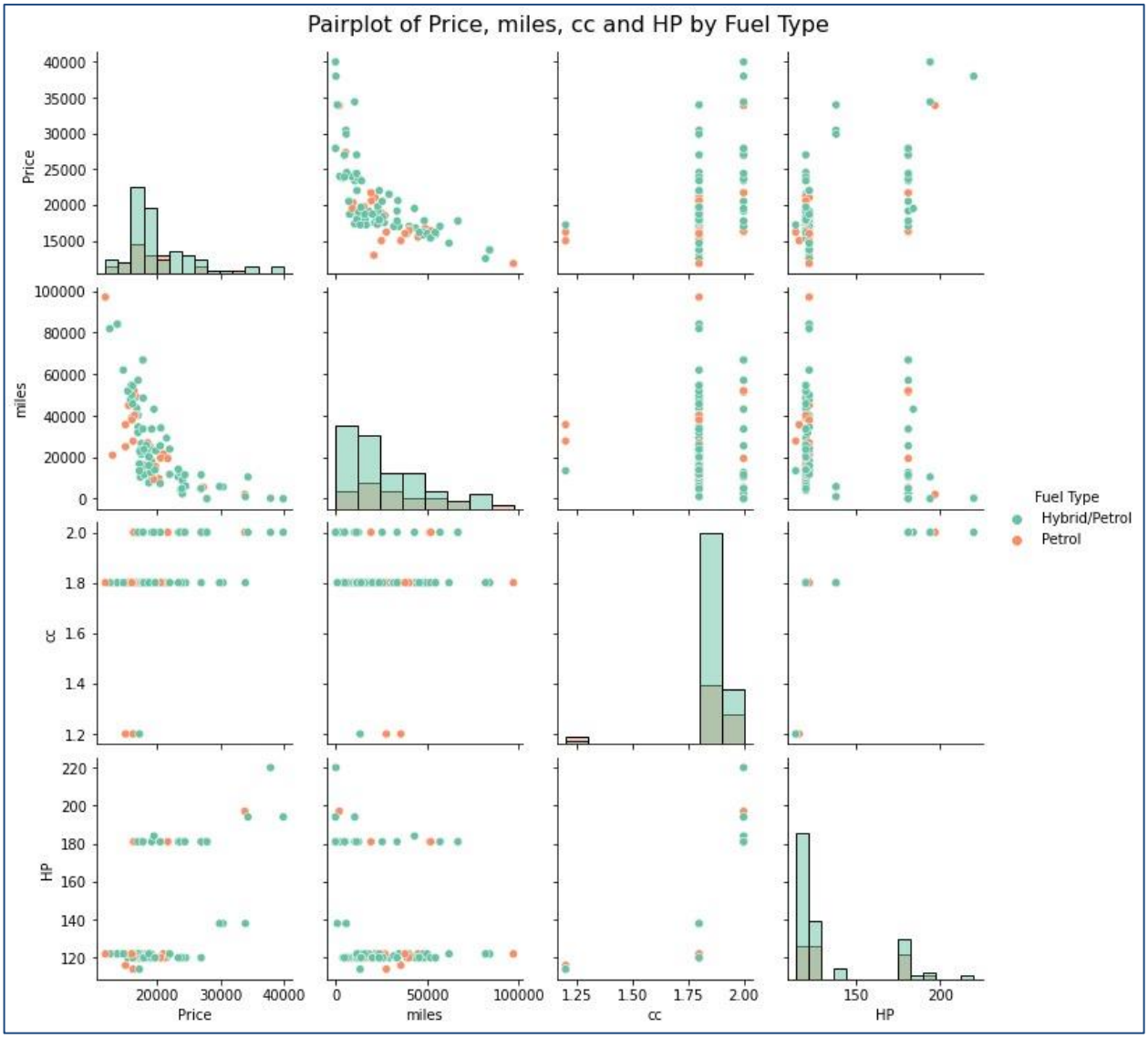
- **Upon pulling nearly 300 population data points—a substantial quantity of data—from the car trader resource website, we noticed that part of the data contained null values, or blank values, which might have an effect on the accuracy of the predictive model. Consequently, we logically shaped the data such that we could appropriately evaluate and analyze the outcomes.**

Aston University
BIRMINGHAM UK

PART 1 - CREATING VISUALIZATIONUALISATION

[DATA VISUALIZATION]

The graphs shown here are a pair plot of four variables: The **pair-plot** provides a comprehensive view of the relationships between vehicle price, miles, engine capacity (cc), and horsepower (HP) for different fuel types. Key insights include a strong negative correlation between price and miles, a weak positive correlation between price and horsepower, and specific groupings between engine capacity and horsepower. Hybrid/Petrol vehicles tend to span a wider range of prices and miles compared to Petrol vehicles, which are more clustered at lower values. The analysis supports the understanding of how these variables interact and vary by fuel type, providing valuable insights for predictive modeling and decision-making in the automotive market

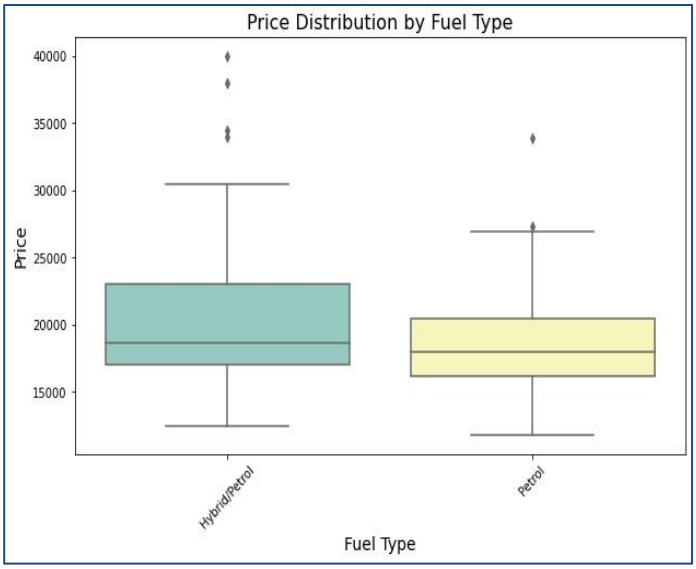


1. **Diagonals (Histograms)**:
    - Each diagonal subplot shows the distribution of a single variable.
    - For example, the first plot on the diagonal shows the distribution of Price, the second shows Miles, the third shows cc, and the fourth shows HP.
    - The colour indicates the Fuel Type, with green representing Hybrid/Petrol and orange representing Petrol.
2. **Off-diagonals (Scatter Plots)**:
    - The scatter plots show the relationships between each pair of variables.
    - For instance, the scatter plot in the first row and second column shows the relationship between Price and Miles.
    - Each point represents an observation, with the colour again indicating the Fuel Type.
    -
3. **Observations**:
    - **Price vs. Miles**: There is a visible negative relationship, meaning as the miles increase, the price tends to decrease. This pattern is consistent for both Hybrid/Petrol and Petrol vehicles.
    - **Price vs. cc**: The prices are clustered at certain cc values, indicating that most cars have specific engine sizes.
    - **Price vs. HP**: There's no clear pattern, but it shows the distribution of prices across different horsepower values.
    - **Miles vs. cc and HP**: No clear relationship is visible, indicating that mileage doesn't have a straightforward correlation with engine size (cc) or horsepower (HP).
    - **cc vs. HP**: This plot shows that most cars have similar cc values with varying horsepower, with both fuel types intermixed.

4. **Fuel Type Comparison**:
    - The pair plot allows you to compare the distribution and relationships of these variables for different fuel types.
    - For example, you can observe if Hybrid/Petrol cars have different mileage distributions compared to Petrol cars or if their prices tend to be higher or lower.

In summary, the pair plot is a powerful tool for exploring the relationships and distributions of multiple variables simultaneously, providing insights into potential correlations and differences between categories (Fuel Types in this case).

(Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.)



The **box plot** displayed is a visualization of the price distribution of vehicles based on their fuel type.

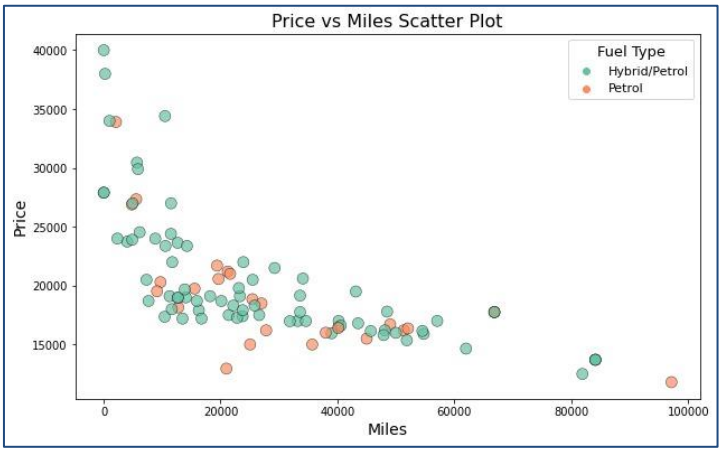
The graph shows that vehicles using Hybrid/Petrol tend to have higher and more variable prices compared to Petrol vehicles. There are also more high-priced outliers among Hybrid/Petrol vehicles, suggesting a wider range of prices for this fuel type

The **whiskers** reach the minimum and greatest values within 1.5 times the interquartile range from the quartiles.

Outliers are represented by individual dots and indicate prices that are significantly higher than the rest of the data.

The **Scatter plot** displayed is a visualization of the relationship between vehicle price and the number of miles driven, differentiated by fuel type. Here is an explanation of the components and insights from the graph:

- Each point on the scatter plot represents a vehicle.
- The colour of the points indicates the fuel type: **Green Dots**: Hybrid/Petrol vehicles.
  Orange Dots: Petrol vehicles.



Summary: The scatter plot illustrates a clear inverse relationship between vehicle price and mileage, meaning vehicles with fewer miles generally cost more. Hybrid/Petrol vehicles show a wider spread in prices across different mileages compared to Petrol vehicles. The plot effectively shows the impact of mileage on vehicle pricing and highlights how this relationship varies between the two fuel types.

We may claim that graphs have the correct dimensions, are labeled to represent the attributes of the car, and have a linear connection with the continuous variables by applying Tufte's concept of graphical integrity.

PART 2. DEPLOYEMENT OF STATISTICAL METHOD AND TEST

[CORRELATION]

The correlation matrix provides a comprehensive view of the relationships between different vehicle-related variables. Price is significantly influenced by HP and Registration, and negatively influenced by miles. The colour blue does not show significant correlations with other variables. This table helps in understanding the underlying relationships and the statistical significance of these relationships, which can be useful for predictive modeling and decision-making in vehicle-related analysis. (Levine, D. M., Stephan, D. F., Krehbiel, T. C., & Berenson, M. L. (2010). *Statistics for Managers Using Microsoft Excel* (6th ed.). Pearson Education.)

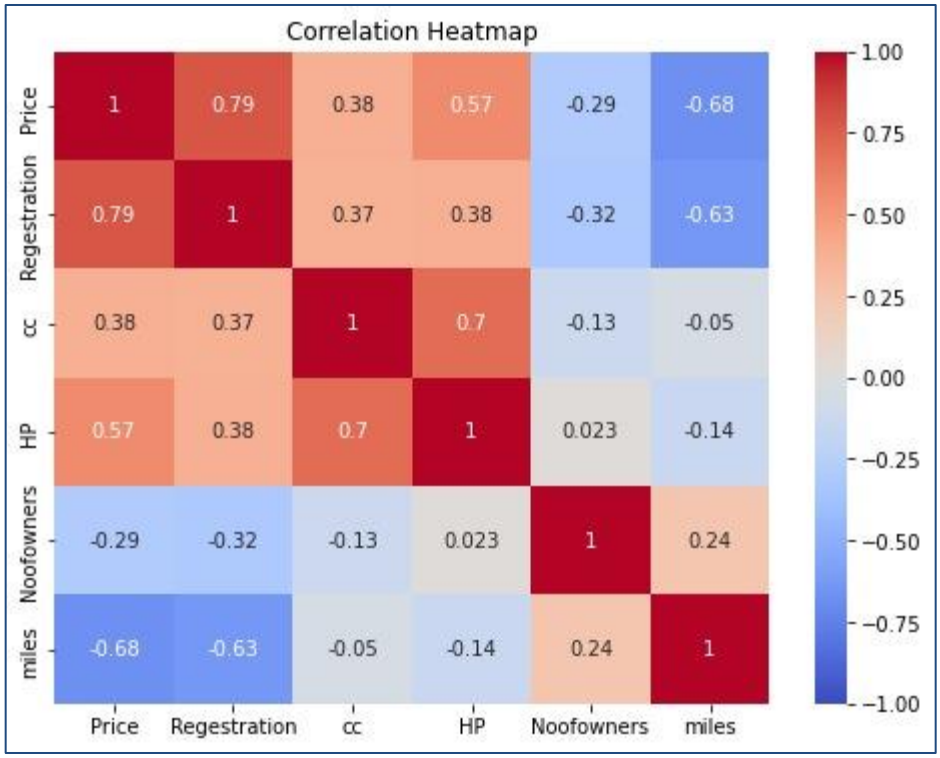
INSIGHTS

- Positively correlated with HP (0.569) and Registration (0.786), both correlations are statistically significant (p-value < 0.001).
- Negatively correlated with miles (-0.677), indicating that as mileage increases, the price tends to decrease, which is statistically significant (p-value < 0.001).
- Weak positive correlation with Colour=Blue (0.060), not statistically significant (p-value = 0.275).

Correlations

		Price	Colour=Blue	miles	HP	Regestration
Pearson Correlation	Price	1.000	.060	-.677	.569	.786
	Colour=Blue	.060	1.000	.053	-.075	-.045
	miles	-.677	.053	1.000	-.142	-.634
	HP	.569	-.075	-.142	1.000	.380
	Regestration	.786	-.045	-.634	.380	1.000
Sig. (1-tailed)	Price	.	.275	<.001	<.001	<.001
	Colour=Blue	.275	.	.301	.229	.327
	miles	.000	.301	.	.080	.000
	HP	.000	.229	.080	.	.000
	Regestration	.000	.327	.000	.000	.
N	Price	100	100	100	100	100
	Colour=Blue	100	100	100	100	100
	miles	100	100	100	100	100
	HP	100	100	100	100	100
	Regestration	100	100	100	100	100

Aston University
BIRMINGHAM UK



The **Heatmap** displayed is a visualization of the correlation matrix between various variables related to vehicles.

Both the x-axis and y-axis represent the same set of variables: Price, Registration, cc (engine capacity), HP (horsepower), Noofowners (number of owners), and miles.

- Highly positively correlated with Price (0.79).
- Moderately positively correlated with cc (0.37) and HP (0.38).
- Negatively correlated with miles (-0.63) and Noofowners (-0.32).

*The colour scale on the right side of the heatmap shows the strength and direction of the correlation.

*Colour Red indicate a positive correlation, with darker reds showing stronger correlations near to +1.

*Blue colours indicate a negative correlation, with darker blues showing stronger correlations near to -1.

*White or light colours indicate weak or no correlation, near to 0.

Summary: The heatmap provides a quick visual summary of the correlations between different vehicle-related variables. It shows that Price is strongly influenced by Registration and HP, and negatively influenced by miles driven. Similarly, Registration shows a strong positive correlation with Price and HP, and a strong negative correlation with miles. Engine capacity (cc) and horsepower (HP) are also closely related. Miles driven is inversely related to both Price and Registration, suggesting that higher mileage tends to lower the vehicle's price and registration value.

Aston University
BIRMINGHAM UK

[CHI-SQUARE TEST]

A Chi-square test is a statistical method used to determine if there is a significant association between two categorical variables. (McHugh, M. L. (2013). The Chi-square test of independence. Biochemia *Medica*, 23(2), 143-149. doi:10.11613/BM.2013.018)

Interpretation: The p-value (0.089) is greater than the common significance level (0.05), indicating that there is no statistically significant association between the variables at the 5% significance level.

- **Value**: 469.980 **Number of cases analyzed**: 100
- **Degrees of Freedom (df)**: 430
- **Asymptotic Significance (2-sided)**: 0.089

RATIO

- **Value**: 317.617
- **Degrees of Freedom (df)**: 430
- **Asymptotic Significance (2-sided)**: 1.000

Interpretation: The p-value (1.000) indicates an extremely low likelihood of a significant association between the variables.

Linear-by-Linear Association:

- **Value**: 61.158
- **Degrees of Freedom (df)**: 1
- **Asymptotic Significance (2-sided)**: <0.001
- Interpretation: The p-value (<0.001) indicates a strong linear association between the variables, suggesting a significant relationship.

Pearson Chi-Square The results of the Likelihood Ratio and Pearson Chi-Square tests, with p-values larger than 0.05, show that there is no significant link between the variables. Put another way, there is insufficient data to conclude that the variables are connected.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	469.980[a]	430	.089
Likelihood Ratio	317.617	430	1.000
Linear-by-Linear Association	61.158	1	<.001
N of Valid Cases	100		

a. 522 cells (100.0%) have expected count less than 5. The minimum expected count is .03.

However, the Linear-by-Linear Association test shows a different result. It suggests that there is a significant linear relationship

between the variables, as its p-value is less than 0.001. This means that as one variable changes, the other variable tends to change in a predictable way.

[DESCRIPTIVE STATISTIC]

The table summarizes the major facts for several elements influencing second-hand automobile market pricing. Here is what each statistic signifies in simple terms.

As we can see the significant information for several variables connected to second-hand automobile market prices.

On average, the automobiles cost roughly £19,937.99, with prices differing by about £5,318.90.

Approximately 25% of the automobiles in the sample are blue, suggesting some variation in this proportion.

Descriptive Statistics

	Mean	Std. Deviation	N
Price	19937.99	5318.895	100
Colour=Blue	.2500	.43519	100
miles	28000.20	21908.859	100
HP	135.78	27.380	100
Regestration	2020.95	1.410	100

The average mileage for the automobiles is 28,000.20 miles, although this figure varies significantly, with a standard variation of 21,908.86 miles.

The vehicles' average horsepower is 135.78, with a variance of 27.38 HP. Finally, the average registration year is 2020, with a minor deviation of 1.41 years.

This table provides a glimpse of the dataset, showing the average values and variances for automobile pricing, colour, mileage, horsepower, and registration year

[ANOVA Table]

The regression model with variables (Colour=Blue, Registration, HP, miles) strongly predicts car pricing, as the ANOVA table demonstrates. The model is trustworthy for this purpose since it explains a sizable amount of the variance in car prices, as shown by the strong F-statistic and extremely low p-value.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2225177791.7	4	556294447.92	91.814	<.001[b]
	Residual	575595959.31	95	6058904.835		
	Total	2800773751.0	99			

a. Dependent Variable: Price

b. Predictors: (Constant), Regestration, Colour=Blue, HP, miles

• **Regression Model Significance**: The F-statistic (91.814) is significantly high, and the p-value (< 0.001) is very low, indicating that the regression model provides a better fit to the data than a model with no predictors.

• **Predictive Power**: The regression model explains a substantial portion of the variability in vehicle prices, as indicated by the high sum of squares for the regression.

Sum of Squares (SS):	**Degrees of Freedom (df)**:	**Mean Square (MS)**:
Regression: 4	**Regression**: 556294447.92	**Regression**: 556294447.92
Residual: 575595959.31	**Residual**: 95	**Residual**: 6058904.835
Total: 2800773751.0	**Total**: 99	

F-Statistic:

- **Value**: 91.814
  - Ratio of the mean square from the regression model to the mean square of the residuals (MS Regression / MS Residual).
- **Significance (Sig.)**: < 0.001

  o The probability that the observed link happened by coincidence is indicated by the p-value. If the number is less than 0.001, it indicates that the model is very significant.

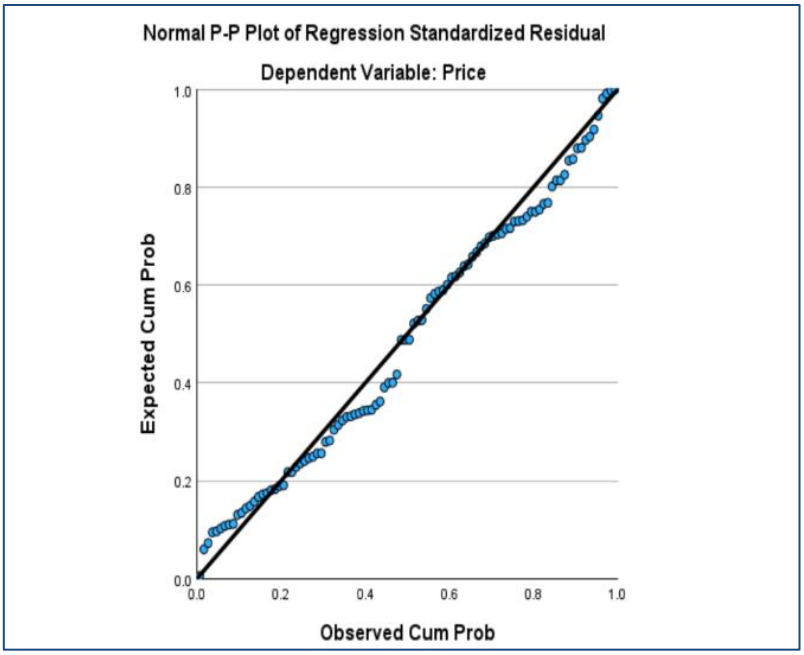
PART 3 - Regression and Residual Analysis

The regression and residual analysis confirm the model's suitability for predicting vehicle prices. Significant predictors (Registration, HP, and miles) strongly influence price. Residual analysis supports normality, linearity, and homoscedasticity, indicating a good model fit. (Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill/Irwin.)

Fit Line: A straight line drawn through the plot.
It represents the cumulative probability that would be expected given that the residuals are normally distributed.

Dots: Represent the actual cumulative probabilities of the residuals.

Distance from Line: If the residuals are normally distributed,
the dots will fall relatively close to the straight line. Any departure from such a line represents a departure from normality.



• A parsimonious model should be extremely straightforward and provide a fair amount of explanation for the data without being overly complex. Parsimony in regression analysis typically includes the following: choosing a model with fewer variables but one that can still account for a significant amount of the variance in the dependent variable—regression analysis's four underlying assumptions: linearity, independence, homoscedasticity, and residual normality

Good Fit: The points form roughly a straight line, which indicates that the residuals may be normally distributed. This is consistent with the assumption of normality.

Potential Deviations: Some deviations from the line are to be expected and are typically small and not concerning. If there are significant deviations in the tails, there may be concerns about normality.

Statistical Significance: The F-statistic is highly significant ($p < 0.001$), confirming that the model is statistically significant.

Error Metrics: The standard error of the estimate (2461.484) indicates the average prediction error.

Insights:

-The average car price in this dataset is around £19,938.

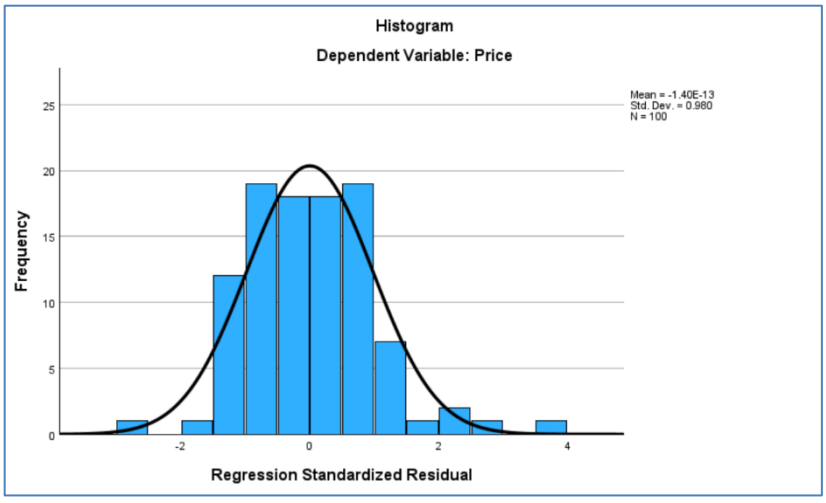
-Roughly 25% of the cars are blue.

-They have an average mileage of around 28,000 miles.

-The average power of the cars is around 136 HP.

-The average year of registration is close to 2021.

These descriptive statistics render an essential general insight into the distribution and central tendencies of the data, which is mandatory for the regression studies one would like to perform further.



The standardized residuals of the regression model used to estimate car price are roughly normally distributed, as the **histogram** demonstrates. The bell-shaped curve illustrates this. The regression model's standardized residuals for estimating car pricing are roughly normally distributed, as the histogram demonstrates. (Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.)

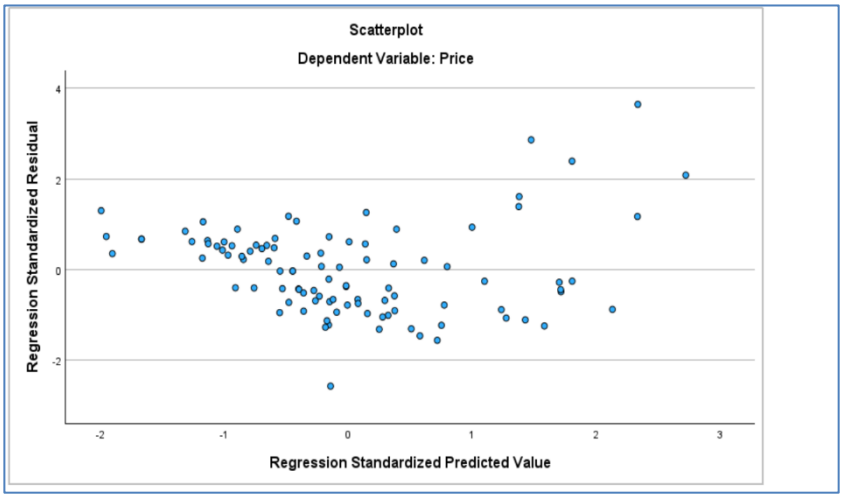
. This is indicated by the bell-shaped curve that overlays the histogram bars and the mean value close to zero. The standard deviation is 0.980, suggesting a reasonable spread of residuals around the mean. The normal distribution of residuals supports the assumption of normality in regression analysis, implying that the model is appropriate for the data.

- **X-Axis (Regression Standardized Residual)**: Represents the standardized residuals from the regression model.
- **Y-Axis (Frequency)**: Represents the frequency of the standardized residuals in each bin.

- **Mean**: -1.40E-13 (essentially zero, indicating the residuals are centered around zero).
- **Standard Deviation (Std. Dev.)**: 0.980, indicating the spread of the residuals.

- The histogram appears to be approximately normally distributed, centered around zero.
- Most residuals fall between -2 and 2, with fewer residuals beyond these points, which is typical for normally distributed data.

The **scatterplot** of standardized residuals versus standardized predicted values is used to check the assumptions of the regression model, including linearity, homoscedasticity, and independence of residuals. The random distribution of residuals around zero in this plot suggests that these assumptions are reasonably met for the regression model predicting vehicle prices. The presence of some extreme residuals indicates potential outliers that might need further investigation. Overall, the plot supports the appropriateness of the regression model for the data.



- **X-Axis (Regression Standardized Predicted Value)**: shows the standardized projected values from the regression model
- **Y-Axis (Regression Standardized Residual)**: represents the standardized residuals, or errors, from the regression model.

Dependent Variable: Price - indicates that the residuals and predicted values are related to the regression model predicting vehicle price.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.891[a]	.794	.786	2461.484	.794	91.814	4	95	<.001	2.172

a. Predictors: (Constant), Colour=Blue, Regestration, HP, miles
b. Dependent Variable: Price

As Adjusted R square is 0.786, which is slightly lower than the R square, indicating a small adjustment for the number of predictors.

R Square (Coefficient of Determination)

- **Value**: 0.794
- **Interpretation**: 79.4% of the variance in vehicle prices is explained by the model's predictors (Colour=Blue, Registration, HP, miles).

Adjusted R square

- **Value**: 0.786
- **Interpretation**: Adjusted for the number of predictors in the model, indicating a slightly lower but still strong fit.

Standard Error of the Estimate

- **Value**: 2461.484
- **Interpretation**: the typical separation between the observed values and the regression line. A better match is indicated by smaller values.

F-Statistic (F Change)

- **Value**: 91.814
- **Degrees of Freedom (df1)**: 4 (number of predictors)
- **Degrees of Freedom (df2)**: 95 (sample size minus number of predictors minus one)
- **Significance (Sig. F Change)**: < 0.001
- **Interpretation**: The F-statistic is highly significant, indicating that the regression model provides a better fit to the data than a model with no predictors.

Durbin-Watson Statistic

- **Value**: 2.172
- **Interpretation**: This value is close to 2, suggesting that there is no significant autocorrelation in the residuals.

As per the regression model it is robust and effectively explains the majority of the variation in vehicle prices, with significant predictors and no major issues with residual autocorrelation.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	10496.19	32856.41	19937.99	4740.943	100
Residual	-6320.245	8972.518	.000	2411.245	100
Std. Predicted Value	-1.992	2.725	.000	1.000	100
Std. Residual	-2.568	3.645	.000	.980	100

a. Dependent Variable: Price

The residuals statistics table confirms that the regression model's predictions for vehicle prices are unbiased, with residuals centered around zero. The spread of residuals and predicted values, as indicated by their standard deviations, suggests variability in the data. The standardized residuals are within an acceptable range, supporting the model's adequacy for predicting vehicle prices.

Predicted Values

- The predicted prices range from approximately 10,496 to 32,856, with an average predicted price of 19,938.
- The standard deviation of the predicted values is 4740.943, indicating variability in the predicted prices.

• **Residuals**:

- The residuals range from -6320.245 to 8972.518, with a mean of zero, indicating that the regression model's predictions are unbiased on average.
- The standard deviation of the residuals is 2411.245, indicating the typical deviation of the actual prices from the predicted prices.

• **Standardized Predicted Values**:

- • To make comparisons easier, these numbers are scaled to have a zero mean and a one standard deviation.
- They range from -1.992 to 2.725.

• **Standardized Residuals**:

- The scale used to these variables results in a zero mean and a one standard deviation, making comparison easier.
- They range from -2.568 to 3.645, indicating that most residuals fall within three standard deviations from the mean.

[PROPOSED CAR PRICE CALCULATOR]

Coefficient Table

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound	Correlations Zero-order	Correlations Partial	Correlations Part	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-3191847.777	495543.757		-6.441	<.001	-4175626.518	-2208069.036					
	HP	71.297	9.889	.367	7.210	<.001	51.665	90.929	.569	.595	.335	.835	1.198
	Regestration	1585.494	245.326	.420	6.463	<.001	1098.461	2072.528	.786	.553	.301	.512	1.954
	miles	-.089	.015	-.366	-6.014	<.001	-.118	-.059	-.677	-.525	-.280	.585	1.708
	Colour=Blue	1543.535	570.624	.126	2.705	.008	410.702	2676.367	.060	.267	.126	.992	1.008

a. Dependent Variable: Price

- Constant: -3191847.777
- HP (Horse Power): 71.297
- Registration: 1585.494
- Miles: -0.089
- Colour=Blue: 1543.535

Regression Equation:

$$Price = 3191847.777 + 71.297 \cdot HP + 1585.494 \cdot Registration - 0.089 \cdot Miles + 1543.535 \cdot Colour=Blue$$

Using the regression equation:

$$Price = -3191847.777 + (71.297 \cdot 200) + (1585.494 \cdot 5) - (0.089 \cdot 50000) + 1543.535$$

$$Price = -3191847.777 + 14259.4 + 7927.47 - 4450 + 1543.535$$

$$Price = -3191847.777 + 18280.405$$

$$Price = -3173567.372$$

Given the coefficients from the regression model, you can use the provided equation to calculate the estimated price of a second-hand car based on its horsepower, registration years, mileage, and colour.

Conclusion

This analysis of the second-hand Toyota C-HR market from 2019 to 2024, using data from AutoTrader, highlights key factors affecting car prices. Descriptive statistics show an average price of £19,937.99, with significant variability. Around 25% of the cars are blue, with an average mileage of 28,000.20 miles, horsepower of 135.78 HP, and a typical registration year of 2021.

Correlation analysis reveals that price is significantly influenced by horsepower and registration year, while negatively affected by miles. The Chi-square tests mostly indicate no significant association between categorical variables, except for a strong linear relationship in one case.

The regression model confirms that horsepower, registration year, and miles are significant predictors of car prices, with residuals supporting a good model fit. A proposed car price calculator derived from this model provides a practical tool for estimating car values.

Overall, this study offers valuable insights into the second-hand car market, aiding better decision-making for buyers and sellers by enhancing understanding of price determinants.

References

(Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.)

(Levine, D. M., Stephan, D. F., Krehbiel, T. C., & Berenson, M. L. (2010). *Statistics for Managers Using Microsoft Excel* (6th ed.). Pearson Education.)

(McHugh, M. L. (2013). The Chi-square test of independence. Biochemia *Medica*, 23(2), 143-149. doi:10.11613/BM.2013.018)

(Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.)

(Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill/Irwin.)