Seneca

Academic Year	**2023**
Semester	☒ Fall ☐ Winter ☐ Summer
Course Code - Name	BAN110 - Data Preparation and Handling
Instructor	Muhammad Rehman Zafar
Assessment	Projects

Student ID	**Student Name**	**Role**
141305227	Mamat Jasseh	

Projects

You are required to choose a project from the list of the projects specified in this document and complete it within groups of **three**.

Since this is a group project, it is required to be done in groups of **3**. Each group should have a Group Lead who would be responsible for submitting the project on Blackboard (Please note that not all the members of the group are required to submit the project separately on Blackboard. **One submission from the Group Lead would be sufficient**).

The detailed requirements for each project are available in this document, so please go through the details and fulfil all the requirements to avoid missing any marks.

Finally, follow the below mentioned instructions carefully.

Instructions:

To obtain maximum marks in this assessment, please ensure the followings:

- Don't forget to write your name and ID on the first page of this document. The student IDs and names of all the students in the group should be mentioned along with the roles.
- Submit the project by writing your solution in this document under the Solution heading below. Do not use a separate document. Everything related to the project should be included in this document, e.g., code, screenshots etc.
- This project has a weightage of **24%** marks of the course.
- This is a group project so **only 1 submission from the group lead is required.**
- Group Leads are required to submit the project on Blackboard as instructed. Submissions through email will not be accepted.
- The project deadline is **midnight December 5, 2023**. Submissions after the deadline will not be accepted.
- A separate session for presentation and QA for the project will be scheduled.
- Upload presentation slides separately to the Blackboard.

Rubric:

Your assessment will be graded based on the following rubric:

	Excellent (7 - 10)	**Average (4 – 6.9)**	**Poor (<4)**
Project Completion and Code **(14)**	The project was completed without any errors and output is as expected. Fulfills all/most of the requirements for the project.	The project was completed with few errors. Fulfills some of the requirements for the project.	The project is incomplete. Does not fulfill all/most of the requirements.
Presentation and QA **(5)**	The student has a good contribution to the project. Knows the ins and outs of the project. The student has presented his/her part of the project very well. Knows everything / most of his/her part.	The student has an average contribution to the project. Does not know the whole project. The student has averagely presented his/her part of the project. Knows few of the things about his/her part.	The student has no contribution to the project. Does not know anything / most about the project. The student has poorly presented the project. Does not know much about the project.
Report **(5)**	Student has contributed well in preparing the project report and knows all the aspects of the report.	Student has contributed partially in preparing the project report and knows some aspects of the report.	Student has not contributed in preparing the report.

Project Instructions

You are provided with a few datasets however; you are free to pick any dataset you like to work on as a group. You are required to demonstrate at least the following skills in the project:

1. Dataset and task description
2. Data Import
    - This phase requires you to import the data from the provided excel file into SAS using Proc Import.
3. Dataset Characteristics and Cleaning
    - This phase requires you to clean your data before data analysis phase. You should use at least following concepts to complete this phase:
        1. Extract relevant data from the original dataset
        2. Convert a numeric column to character column or vice versa
        3. Create a new column based on existing columns and use it in your analysis
        4. Identify missing values and remove / replace using an appropriate technique
        5. Use built-in SAS function(s) to perform data cleaning, e.g., extracting year from the data column etc.

For example, if:

- Target variable
    1. If categorical, show the frequency distribution of each of the possible values. Interpret. Is the dataset balanced? Any other comment?
    2. If numerical, show the statistics (min, max, mean) and the shape of the distribution of the target variable through a histogram. In some case, numerical target variables need transformation to make data modeling possible.
- Categorical variables
    1. Check and correct errors when necessary.
    2. Check and treat missing values through imputation with the mode.
    3. Create one or more derived variables. Justify why the derived variable is created? Does it answer a specific question? Does it serve for data modeling? Etc..
- Numerical variables
    1. Check (range of values/ less than/larger than) and correct errors by deletion.
    2. Check for missing values and correct through imputation with the mean.
    3. Check the distribution of one or more numerical variables to decide which method to use for outlier detection.
    4. Detect and remove outliers.
    5. Test for normality and plot histogram and QQ plots for a variable with a skewed distribution. Apply a transformation and test for normality again with histogram and QQ plot.

4. Data Analysis
    - This phase requires you to analyze your cleaned dataset to answer at least 3 valid business questions. You are free to pick any business questions you like, however, please keep in mind that picking good business questions to answer would result in better marks.
5. Project Report
    - This phase requires you to create a report in MS Word with the following requirements:
        1. Explain each and every phase of the project (from Phase 1 to 4) along with the screenshots of the output and the related SAS code
        2. Include answers to questions in Phase 4 in your report
        3. Create at least 1 graph / chart in your report which can be simply a Box plot to identify outliers etc.
        4. Make sure not to miss any phase and output of its screenshot

Dataset Options

1. Auto-mpg dataset:
   https://www.kaggle.com/uciml/autompg-dataset
2. Heart disease dataset
   https://www.kaggle.com/ronitf/heart-disease-uci
3. Census income dataset
   https://www.kaggle.com/uciml/adult-census-income
4. Bike sharing dataset
   https://www.kaggle.com/marklvl/bike-sharing-dataset
5. Suicide rates dataset:
   https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
6. Breast Cancer
   https://archive.ics.uci.edu/dataset/14/breast+cancer

You are free to use any other dataset from the following sources. Please make sure the dataset meets the requirements listed in dataset requirements section.

Kaggle: https://www.kaggle.com/datasets
UCI: https://archive.ics.uci.edu/dataset

Solution

1. Dataset and task description

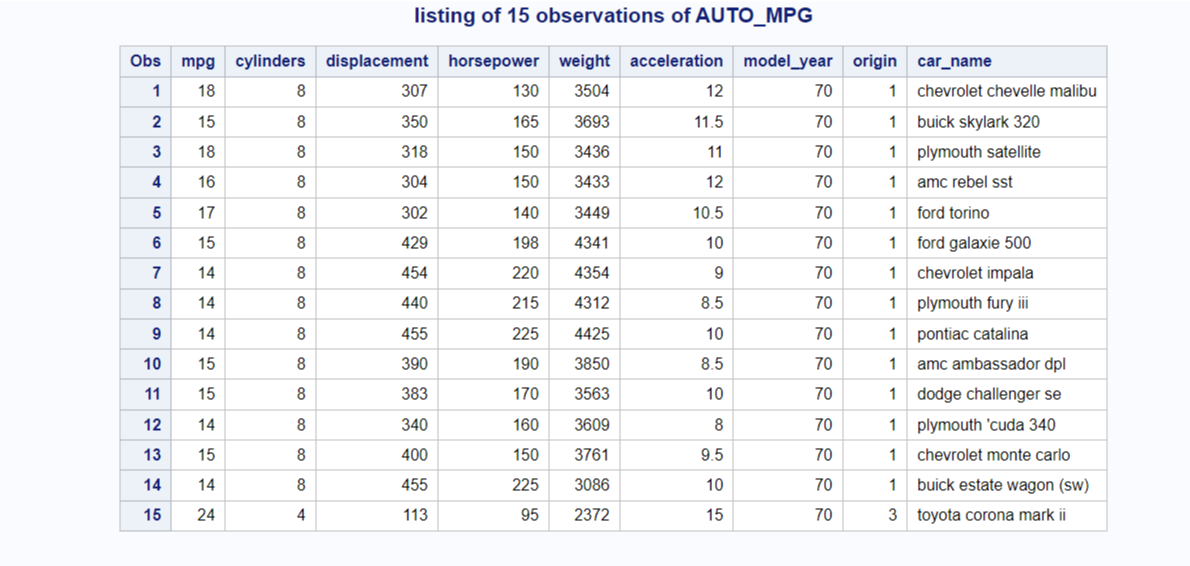
The dataset, Auto_MPG, was chosen for this project, it’s stored in a SAS data library and consists of 398 observations (rows) and nine variables (columns). The variables include a mix of numerical data well as categorical data. This dataset helps analyze automotive attributes related to performance and efficiency, such as determining factors that influence a vehicle's fuel economy or the relationship between a car's weight and its acceleration.

2.Data Import

We used the PROC IMPORT procedure to read external data into SAS. The SAS script imports a CSV file named auto-mpg.csv from a specified directory, creates an SAS dataset named Auto_MPG, and then prints the first 15 observations of this dataset. The getnames=yes option ensures that variable names in the SAS dataset are taken from the first row of the CSV file.

We used PROC CONTENTS to display detailed information about the dataset's structure, such as the types and attributes of the variables it contains. This is particularly useful for understanding the data schema and preparing for further data analysis or manipulation tasks.

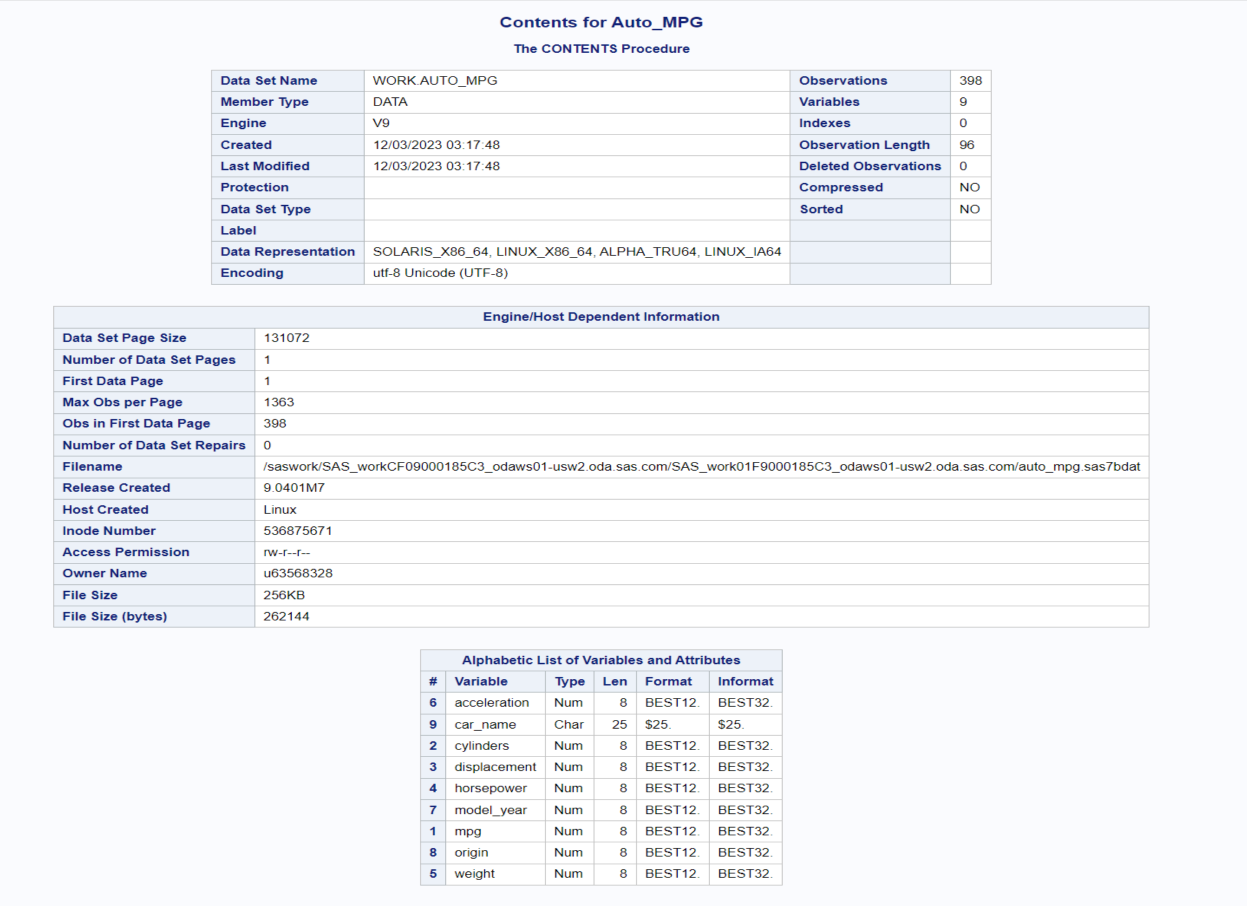
```
/* Q2
loading data*/;
proc import out= Auto_MPG
datafile='/home/u63568328/MY DATA/BAN110/auto-mpg.csv'
DBMS=csv
REPLACE;
getnames=yes;
RUN;

Title "listing of 15 observations of AUTO_MPG";
proc print data=Auto_MPG (obs=15);
run;
```



listing of 15 observations of AUTO_MPG

Obs	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
1	18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
2	15	8	350	165	3693	11.5	70	1	buick skylark 320
3	18	8	318	150	3436	11	70	1	plymouth satellite
4	16	8	304	150	3433	12	70	1	amc rebel sst
5	17	8	302	140	3449	10.5	70	1	ford torino
6	15	8	429	198	4341	10	70	1	ford galaxie 500
7	14	8	454	220	4354	9	70	1	chevrolet impala
8	14	8	440	215	4312	8.5	70	1	plymouth fury iii
9	14	8	455	225	4425	10	70	1	pontiac catalina
10	15	8	390	190	3850	8.5	70	1	amc ambassador dpl
11	15	8	383	170	3563	10	70	1	dodge challenger se
12	14	8	340	160	3609	8	70	1	plymouth 'cuda 340
13	15	8	400	150	3761	9.5	70	1	chevrolet monte carlo
14	14	8	455	225	3086	10	70	1	buick estate wagon (sw)
15	24	4	113	95	2372	15	70	3	toyota corona mark ii

```
*Examining the data-type for dataset;
Title"Contents for Auto_MPG";
proc contents data=Auto_MPG;
run;
```



Contents for Auto_MPG

The CONTENTS Procedure

Data Set Name	WORK.AUTO_MPG	Observations	398
Member Type	DATA	Variables	9
Engine	V9	Indexes	0
Created	12/03/2023 03:17:48	Observation Length	96
Last Modified	12/03/2023 03:17:48	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1363
Obs in First Data Page	398
Number of Data Set Repairs	0
Filename	/saswork/SAS_workCF09000185C3_odaws01-usw2.oda.sas.com/SAS_work01F9000185C3_odaws01-usw2.oda.sas.com/auto_mpg.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	536875671
Access Permission	rw-r--r--
Owner Name	u63568328
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
6	acceleration	Num	8	BEST12.	BEST32.
9	car_name	Char	25	$25.	$25.
2	cylinders	Num	8	BEST12.	BEST32.
3	displacement	Num	8	BEST12.	BEST32.
4	horsepower	Num	8	BEST12.	BEST32.
7	model_year	Num	8	BEST12.	BEST32.
1	mpg	Num	8	BEST12.	BEST32.
8	origin	Num	8	BEST12.	BEST32.
5	weight	Num	8	BEST12.	BEST32.

3. Dataset Characteristics and Cleaning

After importing the data, we performed data manipulation and cleaning tasks on a dataset named auto_mpg. Below is the breakdown and analyze each part of the phase 3:

Selecting Specific Columns:

In this section created a new dataset relevant_data from auto_mpg by keeping only specific columns: mpg, cylinders, horsepower, weight, acceleration, model_year, origin, and car_name.

Converting Origin from Numeric to Character:

A new dataset AutoMPG is created from relevant_data. The origin column, which is numeric, is converted to a character format using the put function. The new character variable is named Char_origin, which is then renamed back to origin, replacing the original numeric column.

Examining Data Type After Conversion:

The proc contents procedure is used to display metadata about the AutoMPG dataset, showing changes in data types and structure.

Creating a New Column for Weight to Horsepower Ratio:

A new column weight_to_horsepower is created in a dataset named mpg, calculated as the ratio of weight to horsepower and rounded to two decimal places.

Identifying Missing Values:

proc freq is used to analyze the horsepower column in the mpg dataset for missing values.

Replacing Missing Values:

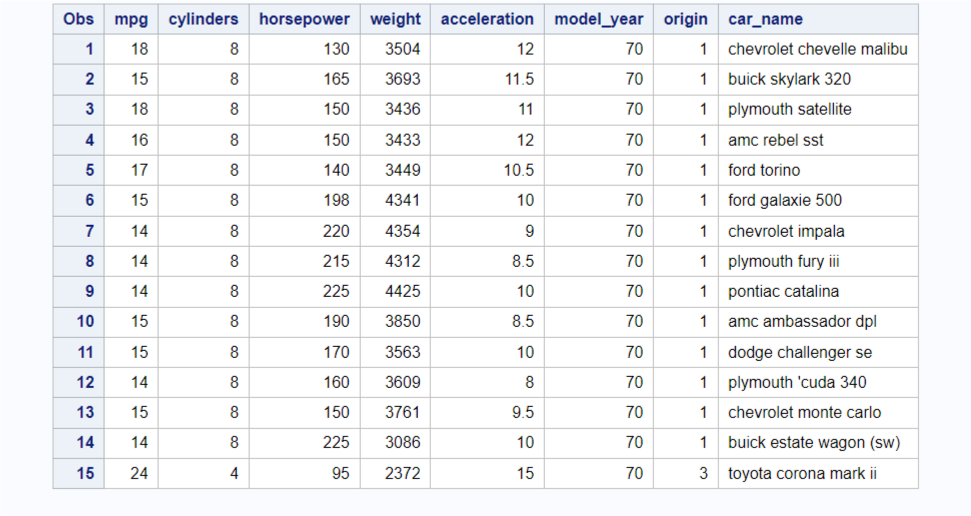
A dataset cleaned_data is created from mpg.

Missing values in the horsepower column denoted by '.' are replaced with '0'. This is an example of handling missing data, but in a real-world scenario, the replacement value should be chosen based on the context and nature of the data.

Extracting a Component from the Car Name:

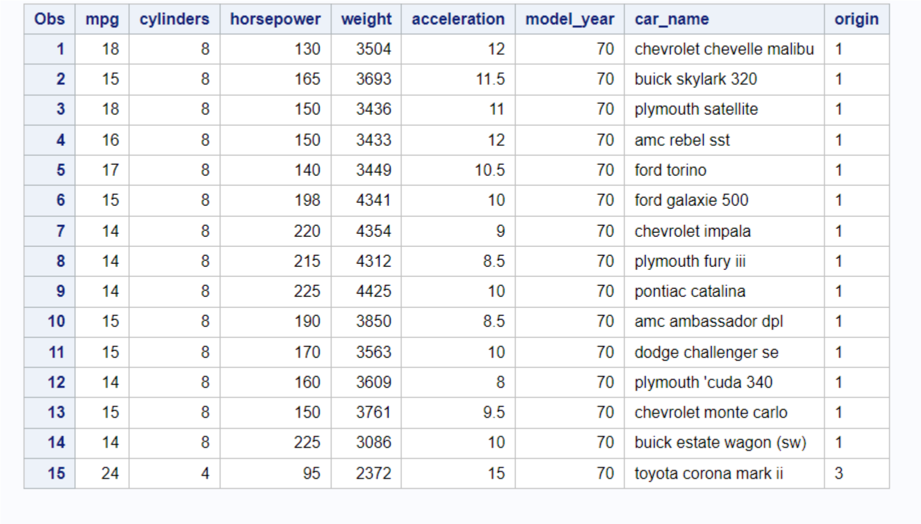
A new column brand is created in the cleaned_data dataset, which extracts the first word from car_name using the scan function. This is assumed to represent the car's brand.

```
*/Q3 Data Characteristics and cleaning*/;

/* Selecting specific columns */
DATA relevant_data;
    SET auto_mpg (keep=mpg cylinders horsepower weight acceleration
model_year origin car_name);
RUN;
proc print data = relevant_data (obs=15);
run;
```



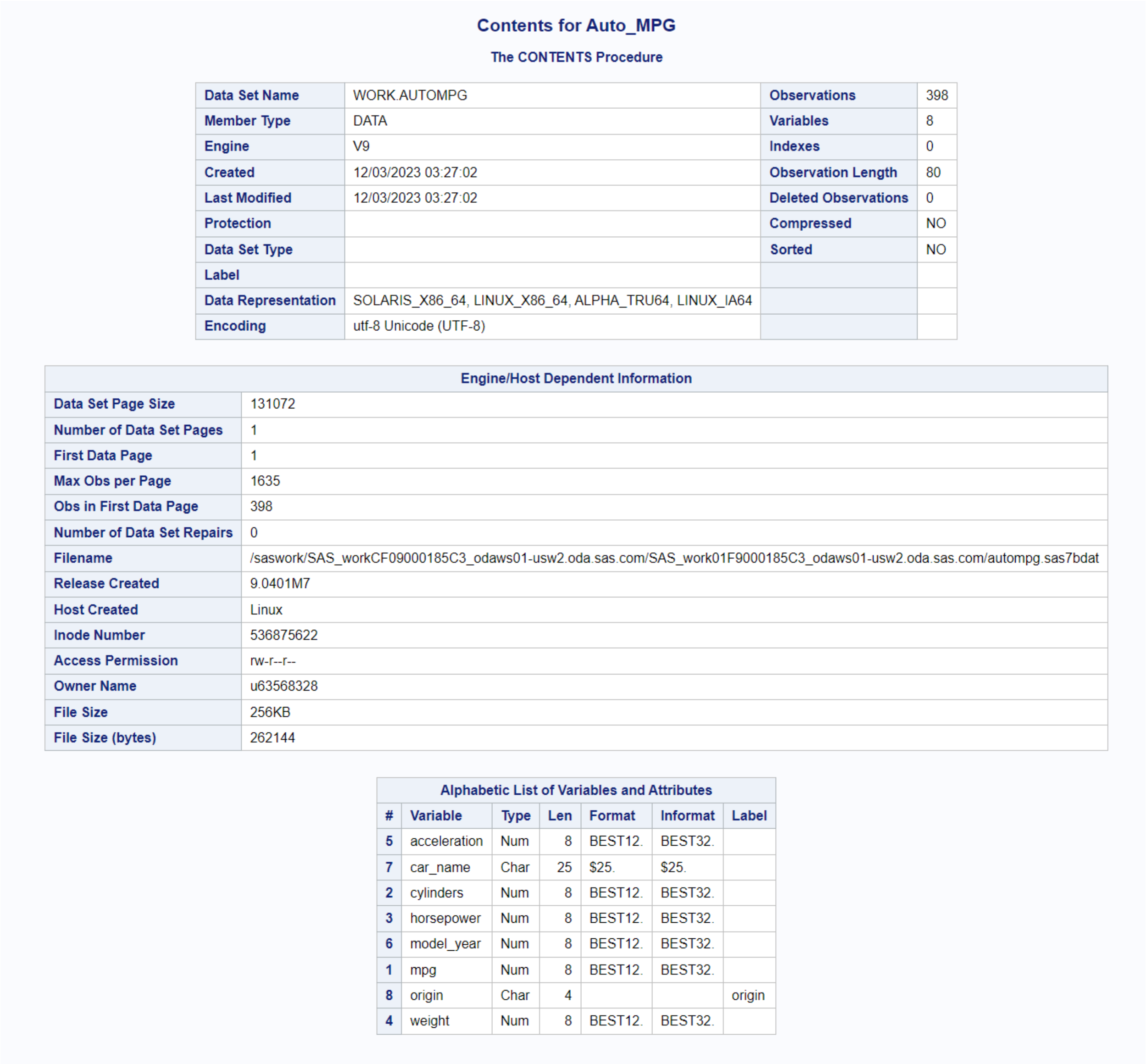
Obs	mpg	cylinders	horsepower	weight	acceleration	model_year	origin	car_name
1	18	8	130	3504	12	70	1	chevrolet chevelle malibu
2	15	8	165	3693	11.5	70	1	buick skylark 320
3	18	8	150	3436	11	70	1	plymouth satellite
4	16	8	150	3433	12	70	1	amc rebel sst
5	17	8	140	3449	10.5	70	1	ford torino
6	15	8	198	4341	10	70	1	ford galaxie 500
7	14	8	220	4354	9	70	1	chevrolet impala
8	14	8	215	4312	8.5	70	1	plymouth fury iii
9	14	8	225	4425	10	70	1	pontiac catalina
10	15	8	190	3850	8.5	70	1	amc ambassador dpl
11	15	8	170	3563	10	70	1	dodge challenger se
12	14	8	160	3609	8	70	1	plymouth 'cuda 340
13	15	8	150	3761	9.5	70	1	chevrolet monte carlo
14	14	8	225	3086	10	70	1	buick estate wagon (sw)
15	24	4	95	2372	15	70	3	toyota corona mark ii

```
/* Convert the 'origin' from numeric to character */
data AutoMPG;
    set relevant_data;
    Char_origin = put(origin, 4.);
    label Char_origin = 'origin';
    drop origin;
    rename Char_origin = origin;
run;
proc print data=autompg (obs=15);
run;
```



Obs	mpg	cylinders	horsepower	weight	acceleration	model_year	car_name	origin
1	18	8	130	3504	12	70	chevrolet chevelle malibu	1
2	15	8	165	3693	11.5	70	buick skylark 320	1
3	18	8	150	3436	11	70	plymouth satellite	1
4	16	8	150	3433	12	70	amc rebel sst	1
5	17	8	140	3449	10.5	70	ford torino	1
6	15	8	198	4341	10	70	ford galaxie 500	1
7	14	8	220	4354	9	70	chevrolet impala	1
8	14	8	215	4312	8.5	70	plymouth fury iii	1
9	14	8	225	4425	10	70	pontiac catalina	1
10	15	8	190	3850	8.5	70	amc ambassador dpl	1
11	15	8	170	3563	10	70	dodge challenger se	1
12	14	8	160	3609	8	70	plymouth 'cuda 340	1
13	15	8	150	3761	9.5	70	chevrolet monte carlo	1
14	14	8	225	3086	10	70	buick estate wagon (sw)	1
15	24	4	95	2372	15	70	toyota corona mark ii	3

```
/*examining the data-type after conversion*/;
Title "Contents for Auto_MPG";
proc contents data=AutoMPG;
run;
```



Contents for Auto_MPG

The CONTENTS Procedure

Data Set Name	WORK.AUTOMPG	**Observations**	398
Member Type	DATA	**Variables**	8
Engine	V9	**Indexes**	0
Created	12/03/2023 03:27:02	**Observation Length**	80
Last Modified	12/03/2023 03:27:02	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information

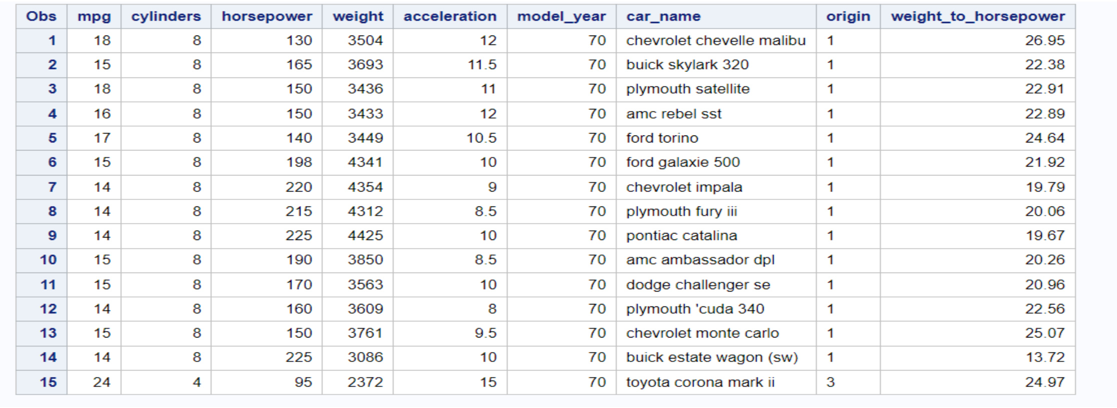
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1635
Obs in First Data Page	398
Number of Data Set Repairs	0
Filename	/saswork/SAS_workCF09000185C3_odaws01-usw2.oda.sas.com/SAS_work01F9000185C3_odaws01-usw2.oda.sas.com/autompg.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	536875622
Access Permission	rw-r--r--
Owner Name	u63568328
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
5	acceleration	Num	8	BEST12.	BEST32.	
7	car_name	Char	25	$25.	$25.	
2	cylinders	Num	8	BEST12.	BEST32.	
3	horsepower	Num	8	BEST12.	BEST32.	
6	model_year	Num	8	BEST12.	BEST32.	
1	mpg	Num	8	BEST12.	BEST32.	
8	origin	Char	4			origin
4	weight	Num	8	BEST12.	BEST32.	

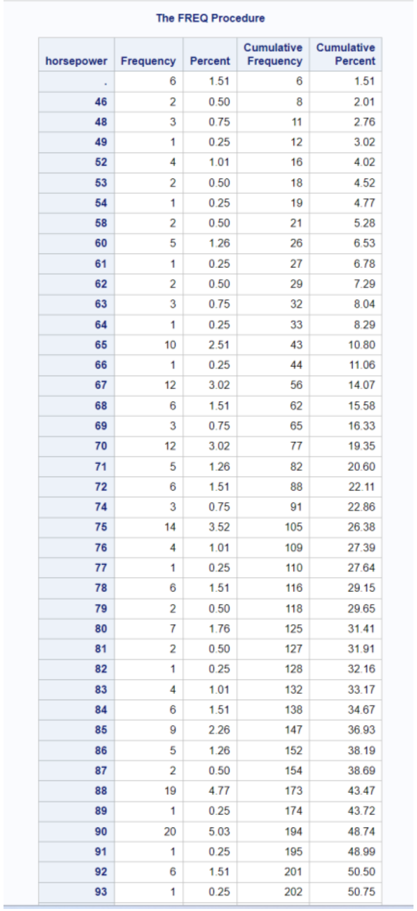
```
/* Creating a new column for weight to horsepower ratio rounded to 2
decimal places */
data mpg;
set autompg;
```

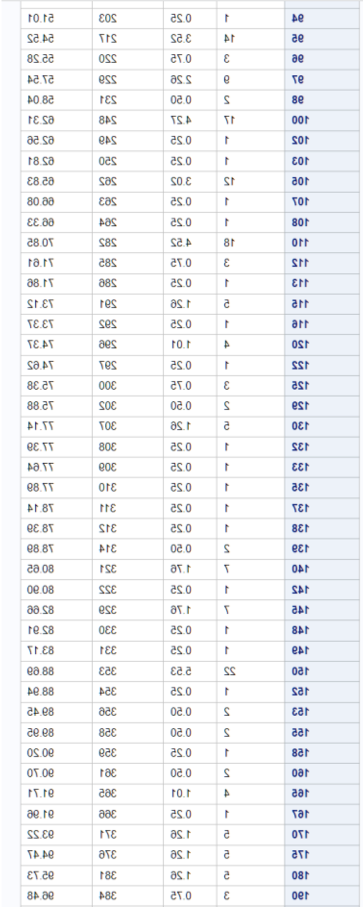
```
weight_to_horsepower = round(weight / horsepower, 0.01);
run;
proc  print data=mpg (obs=15);
run;
```

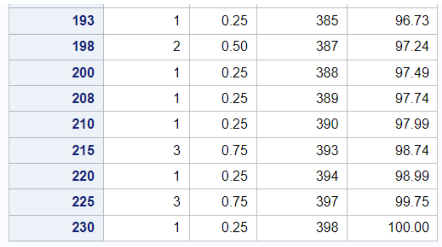


Obs	mpg	cylinders	horsepower	weight	acceleration	model_year	car_name	origin	weight_to_horsepower
1	18	8	130	3504	12	70	chevrolet chevelle malibu	1	26.95
2	15	8	165	3693	11.5	70	buick skylark 320	1	22.38
3	18	8	150	3436	11	70	plymouth satellite	1	22.91
4	16	8	150	3433	12	70	amc rebel sst	1	22.89
5	17	8	140	3449	10.5	70	ford torino	1	24.64
6	15	8	198	4341	10	70	ford galaxie 500	1	21.92
7	14	8	220	4354	9	70	chevrolet impala	1	19.79
8	14	8	215	4312	8.5	70	plymouth fury iii	1	20.06
9	14	8	225	4425	10	70	pontiac catalina	1	19.67
10	15	8	190	3850	8.5	70	amc ambassador dpl	1	20.26
11	15	8	170	3563	10	70	dodge challenger se	1	20.96
12	14	8	160	3609	8	70	plymouth 'cuda 340	1	22.56
13	15	8	150	3761	9.5	70	chevrolet monte carlo	1	25.07
14	14	8	225	3086	10	70	buick estate wagon (sw)	1	13.72
15	24	4	95	2372	15	70	toyota corona mark ii	3	24.97

```
/* Identifying missing values */
proc freq data=mpg;
    tables horsepower / missing;
run;
```



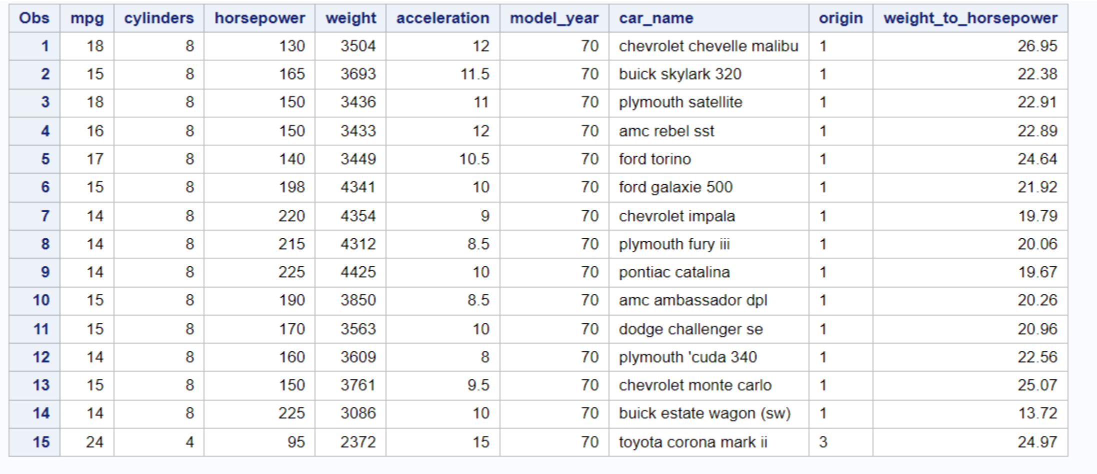




The FREQ Procedure

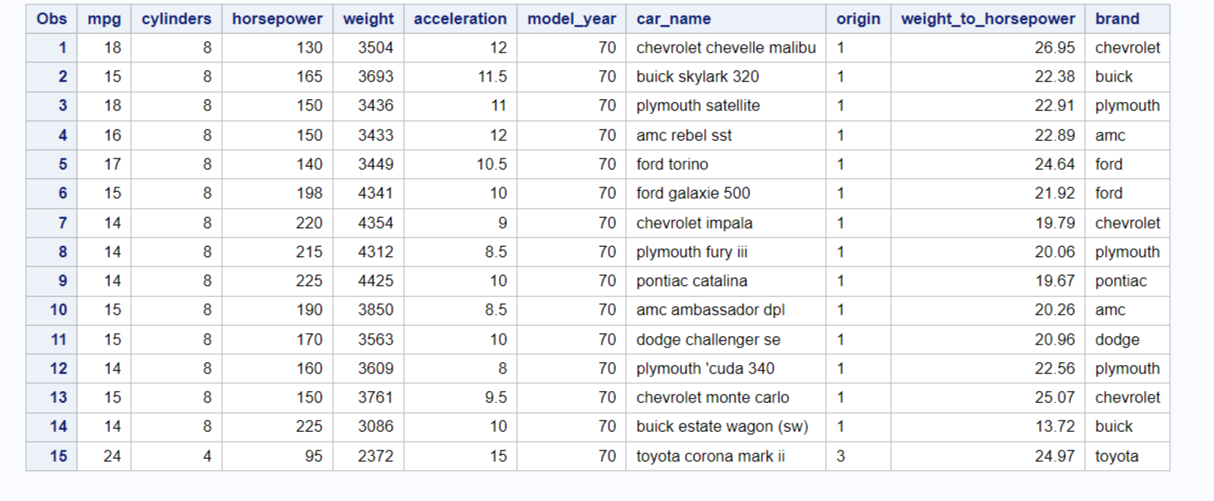
horsepower	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	6	1.51	6	1.51
46	2	0.50	8	2.01
48	3	0.75	11	2.76
49	1	0.25	12	3.02
52	4	1.01	16	4.02
53	2	0.50	18	4.52
54	1	0.25	19	4.77
58	2	0.50	21	5.28
60	5	1.26	26	6.53
61	1	0.25	27	6.78
62	2	0.50	29	7.29
63	3	0.75	32	8.04
64	1	0.25	33	8.29
65	10	2.51	43	10.80
66	1	0.25	44	11.06
67	12	3.02	56	14.07
68	6	1.51	62	15.58
69	3	0.75	65	16.33
70	12	3.02	77	19.35
71	5	1.26	82	20.60
72	6	1.51	88	22.11
74	3	0.75	91	22.86
75	14	3.52	105	26.38
76	4	1.01	109	27.39
77	1	0.25	110	27.64
78	6	1.51	116	29.15
79	2	0.50	118	29.65
80	7	1.76	125	31.41
81	2	0.50	127	31.91
82	1	0.25	128	32.16
83	4	1.01	132	33.17
84	6	1.51	138	34.67
85	9	2.26	147	36.93
86	5	1.26	152	38.19
87	2	0.50	154	38.69
88	19	4.77	173	43.47
89	1	0.25	174	43.72
90	20	5.03	194	48.74
91	1	0.25	195	48.99
92	6	1.51	201	50.50
93	1	0.25	202	50.75
94	1	0.25	203	51.01
95	14	3.52	217	54.52
96	3	0.75	220	55.28
97	9	2.26	229	57.54
98	2	0.50	231	58.04
100	17	4.27	248	62.31
102	1	0.25	249	62.56
103	1	0.25	250	62.81
105	12	3.02	262	65.83
107	1	0.25	263	66.08
108	1	0.25	264	66.33
110	18	4.52	282	70.85
112	3	0.75	285	71.61
113	1	0.25	286	71.86
115	5	1.26	291	73.12
116	1	0.25	292	73.37
120	4	1.01	296	74.37
122	1	0.25	297	74.62
125	3	0.75	300	75.38
129	2	0.50	302	75.88
130	5	1.26	307	77.14
132	1	0.25	308	77.39
133	1	0.25	309	77.64
135	1	0.25	310	77.89
137	1	0.25	311	78.14
138	1	0.25	312	78.39
139	2	0.50	314	78.89
140	7	1.76	321	80.65
142	1	0.25	322	80.90
145	7	1.76	329	82.66
148	1	0.25	330	82.91
149	1	0.25	331	83.17
150	22	5.53	353	88.69
152	1	0.25	354	88.94
153	2	0.50	356	89.45
155	2	0.50	358	89.95
158	1	0.25	359	90.20
160	2	0.50	361	90.70
165	4	1.01	365	91.71
167	1	0.25	366	91.96
170	5	1.26	371	93.22
175	5	1.26	376	94.47
180	5	1.26	381	95.73
190	3	0.75	384	96.48
193	1	0.25	385	96.73
198	2	0.50	387	97.24
200	1	0.25	388	97.49
208	1	0.25	389	97.74
210	1	0.25	390	97.99
215	3	0.75	393	98.74
220	1	0.25	394	98.99
225	3	0.75	397	99.75
230	1	0.25	398	100.00

```
/* Replacing missing values */
data cleaned_data;
    set mpg;
    if horsepower = '.' then horsepower='0'; /* Example replacement */
run;
PROC PRINT DATA=CLEANED_DATA (OBS=15);
RUN;
```



Obs	mpg	cylinders	horsepower	weight	acceleration	model_year	car_name	origin	weight_to_horsepower
1	18	8	130	3504	12	70	chevrolet chevelle malibu	1	26.95
2	15	8	165	3693	11.5	70	buick skylark 320	1	22.38
3	18	8	150	3436	11	70	plymouth satellite	1	22.91
4	16	8	150	3433	12	70	amc rebel sst	1	22.89
5	17	8	140	3449	10.5	70	ford torino	1	24.64
6	15	8	198	4341	10	70	ford galaxie 500	1	21.92
7	14	8	220	4354	9	70	chevrolet impala	1	19.79
8	14	8	215	4312	8.5	70	plymouth fury iii	1	20.06
9	14	8	225	4425	10	70	pontiac catalina	1	19.67
10	15	8	190	3850	8.5	70	amc ambassador dpl	1	20.26
11	15	8	170	3563	10	70	dodge challenger se	1	20.96
12	14	8	160	3609	8	70	plymouth 'cuda 340	1	22.56
13	15	8	150	3761	9.5	70	chevrolet monte carlo	1	25.07
14	14	8	225	3086	10	70	buick estate wagon (sw)	1	13.72
15	24	4	95	2372	15	70	toyota corona mark ii	3	24.97

```
/* Extracting a component from the car name */
data cleaned_data;
    set cleaned_data;
    brand = scan(car_name, 1, ' '); /* Extracts the first word from
car_name */
run;
PROC PRINT DATA=cleaned_data (obs=15);
run;
```



Obs	mpg	cylinders	horsepower	weight	acceleration	model_year	car_name	origin	weight_to_horsepower	brand
1	18	8	130	3504	12	70	chevrolet chevelle malibu	1	26.95	chevrolet
2	15	8	165	3693	11.5	70	buick skylark 320	1	22.38	buick
3	18	8	150	3436	11	70	plymouth satellite	1	22.91	plymouth
4	16	8	150	3433	12	70	amc rebel sst	1	22.89	amc
5	17	8	140	3449	10.5	70	ford torino	1	24.64	ford
6	15	8	198	4341	10	70	ford galaxie 500	1	21.92	ford
7	14	8	220	4354	9	70	chevrolet impala	1	19.79	chevrolet
8	14	8	215	4312	8.5	70	plymouth fury iii	1	20.06	plymouth
9	14	8	225	4425	10	70	pontiac catalina	1	19.67	pontiac
10	15	8	190	3850	8.5	70	amc ambassador dpl	1	20.26	amc
11	15	8	170	3563	10	70	dodge challenger se	1	20.96	dodge
12	14	8	160	3609	8	70	plymouth 'cuda 340	1	22.56	plymouth
13	15	8	150	3761	9.5	70	chevrolet monte carlo	1	25.07	chevrolet
14	14	8	225	3086	10	70	buick estate wagon (sw)	1	13.72	buick
15	24	4	95	2372	15	70	toyota corona mark ii	3	24.97	toyota

4. Data Analysis

Business Questions

1. How does the number of cylinders in a car affect its fuel efficiency (mpg)?
2. Has there been a significant improvement in the horsepower of cars over the years without compromising fuel efficiency?
3. Is there a correlation between the weight of a car and its acceleration?

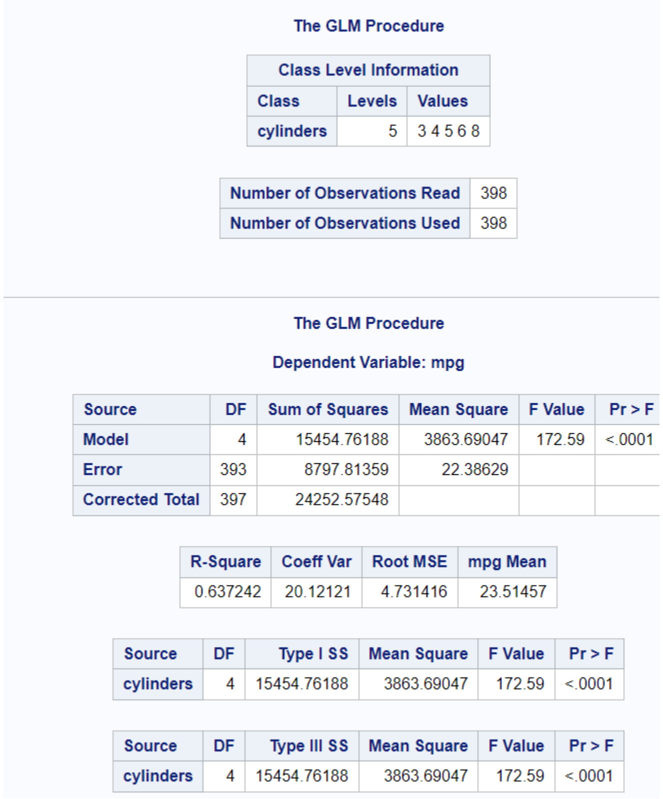
To answer these questions, we performed statistical analyses, including correlation tests, regression analysis, and trend analysis. Here's a brief outline of the SAS procedures we used for each question.

Business Question 1

How does the number of cylinders in a car affect its fuel efficiency (mpg)?

/*Data Analysis*/

```
PROC GLM DATA=cleaned_data;
CLASS cylinders;
MODEL mpg = cylinders;
MEANS cylinders / TUKEY;
RUN;
```



The GLM Procedure

Class Level Information		
Class	Levels	Values
cylinders	5	3 4 5 6 8

Number of Observations Read	398
Number of Observations Used	398

The GLM Procedure

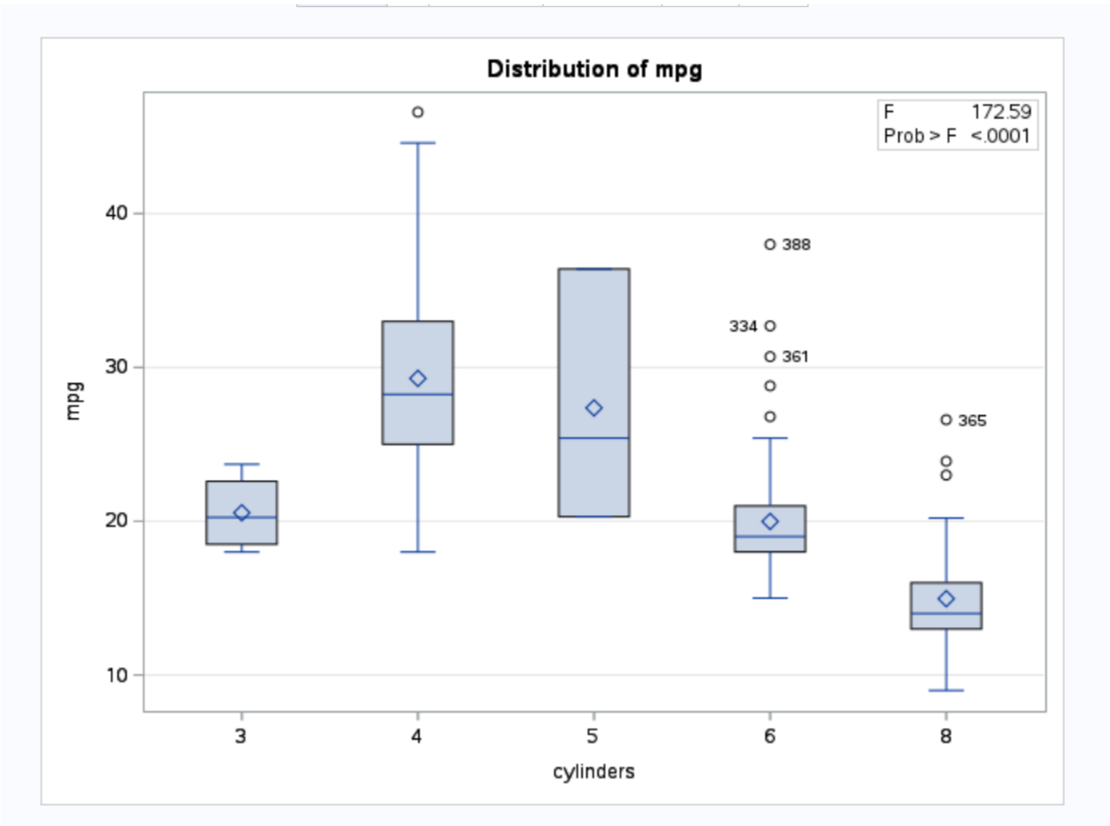
Dependent Variable: mpg

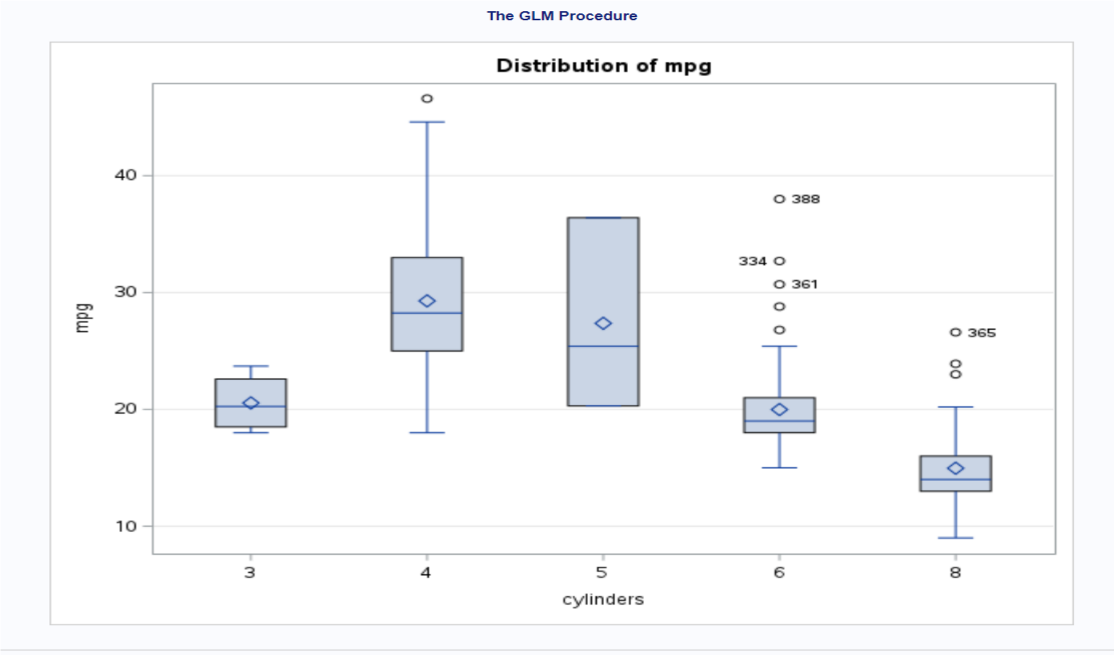
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	15454.76188	3863.69047	172.59	<.0001
Error	393	8797.81359	22.38629		
Corrected Total	397	24252.57548			

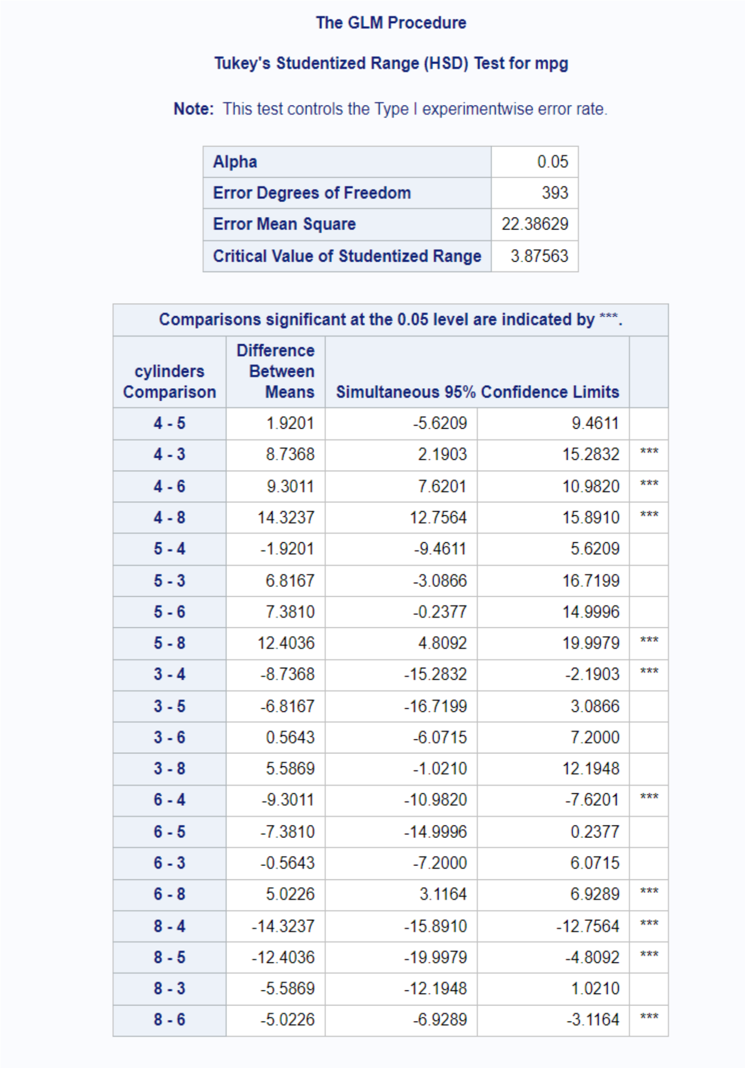
R-Square	Coeff Var	Root MSE	mpg Mean
0.637242	20.12121	4.731416	23.51457

Source	DF	Type I SS	Mean Square	F Value	Pr > F
cylinders	4	15454.76188	3863.69047	172.59	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
cylinders	4	15454.76188	3863.69047	172.59	<.0001




The GLM Procedure



The GLM Procedure

Tukey's Studentized Range (HSD) Test for mpg

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	393
Error Mean Square	22.38629
Critical Value of Studentized Range	3.87563

Comparisons significant at the 0.05 level are indicated by ***.

cylinders Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
4 - 5	1.9201	-5.6209	9.4611	
4 - 3	8.7368	2.1903	15.2832	***
4 - 6	9.3011	7.6201	10.9820	***
4 - 8	14.3237	12.7564	15.8910	***
5 - 4	-1.9201	-9.4611	5.6209	
5 - 3	6.8167	-3.0866	16.7199	
5 - 6	7.3810	-0.2377	14.9996	
5 - 8	12.4036	4.8092	19.9979	***
3 - 4	-8.7368	-15.2832	-2.1903	***
3 - 5	-6.8167	-16.7199	3.0866	
3 - 6	0.5643	-6.0715	7.2000	
3 - 8	5.5869	-1.0210	12.1948	
6 - 4	-9.3011	-10.9820	-7.6201	***
6 - 5	-7.3810	-14.9996	0.2377	
6 - 3	-0.5643	-7.2000	6.0715	
6 - 8	5.0226	3.1164	6.9289	***
8 - 4	-14.3237	-15.8910	-12.7564	***
8 - 5	-12.4036	-19.9979	-4.8092	***
8 - 3	-5.5869	-12.1948	1.0210	
8 - 6	-5.0226	-6.9289	-3.1164	***

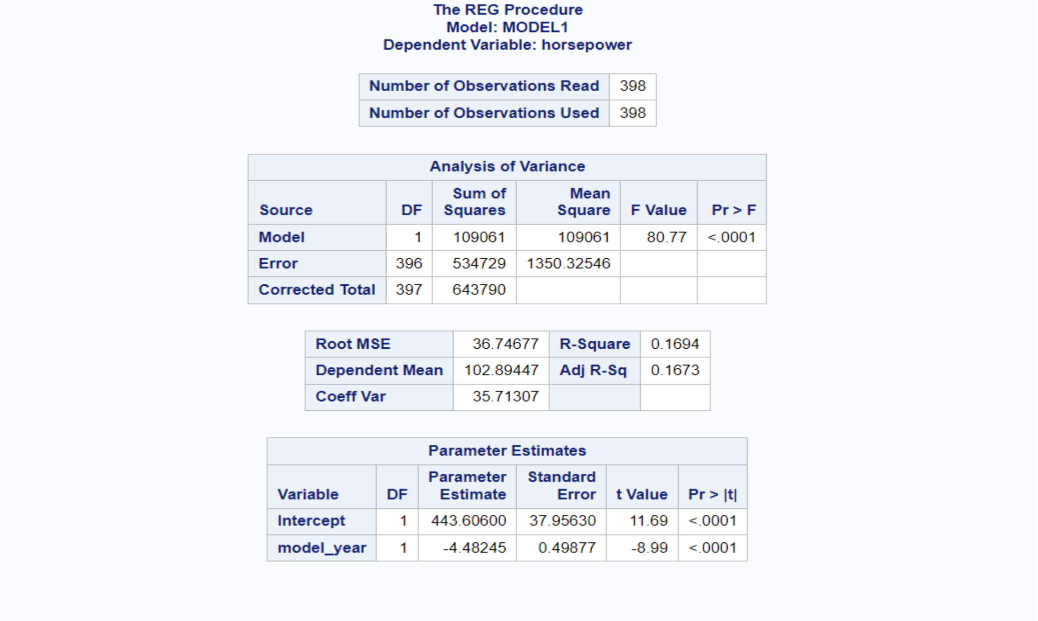
Answer to the Business Question 1:

The number of cylinders in a car has a notable impact on its fuel efficiency, with a general trend of decreased efficiency as the number of cylinders increases. According to Tukey's test results, cars with 4 cylinders are significantly more fuel-efficient than those with 6 or 8 cylinders. While cars with 3 cylinders are also more fuel-efficient than their 4-cylinder counterparts, the sample size may be smaller (as indicated by the degrees of freedom). Moving from 4 to 6 and 6 to 8 cylinders results in a significant decrease in mpg. Overall, for those prioritizing fuel efficiency, choosing a car with fewer cylinders is the better option. However, it's important to note that while cylinder count is a strong predictor of mpg, other factors can also come into play and should be considered when deciding on a vehicle purchase or design.

Business Question 2

Has there been a significant improvement in the horsepower of cars over the years without compromising fuel efficiency?

```
PROC REG DATA=cleaned_data;
MODEL horsepower = model_year;
PLOT horsepower*model_year;
RUN;
```



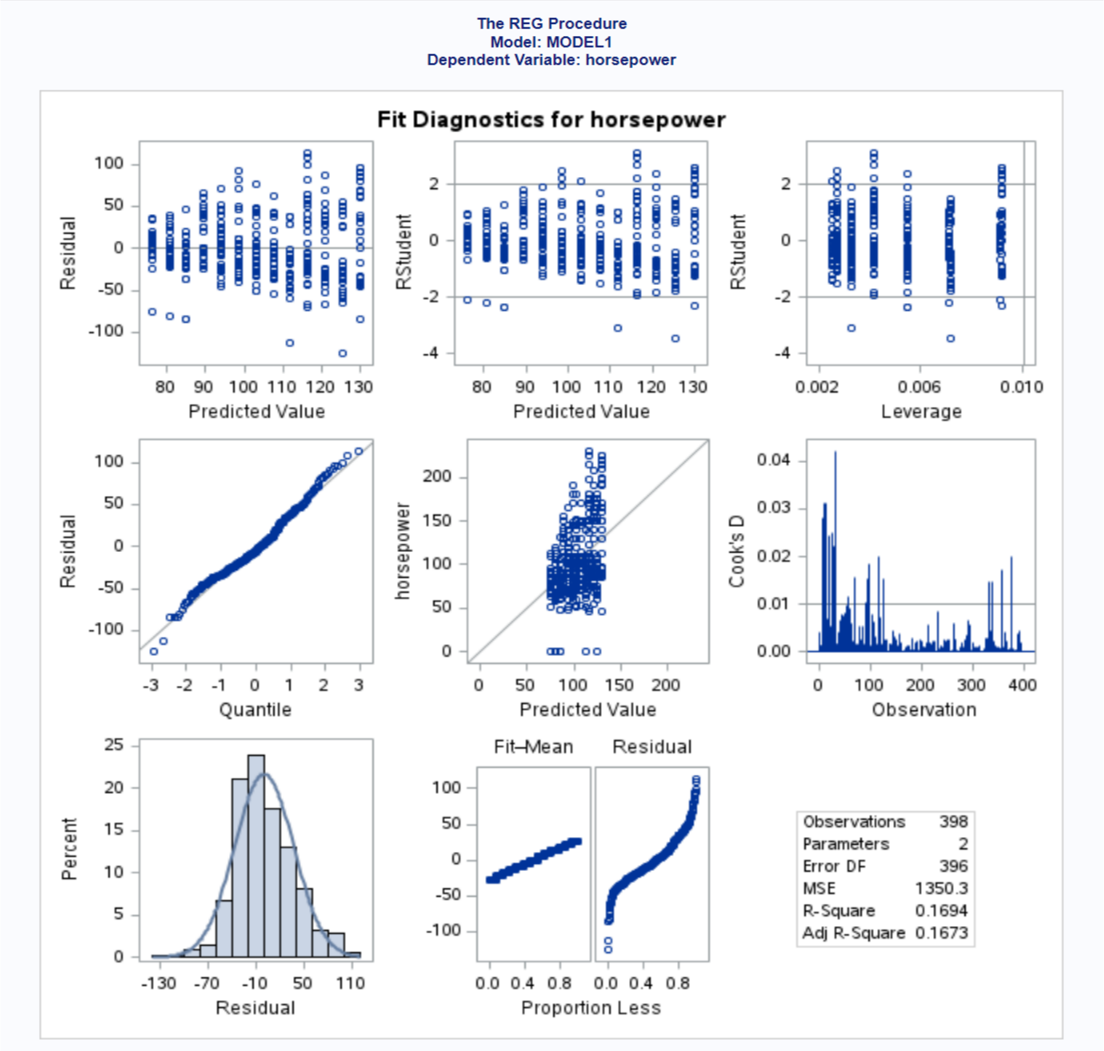
The REG Procedure
Model: MODEL1
Dependent Variable: horsepower

Number of Observations Read	398
Number of Observations Used	398

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	109061	109061	80.77	<.0001
Error	396	534729	1350.32546		
Corrected Total	397	643790			

Root MSE	36.74677	R-Square	0.1694
Dependent Mean	102.89447	Adj R-Sq	0.1673
Coeff Var	35.71307		

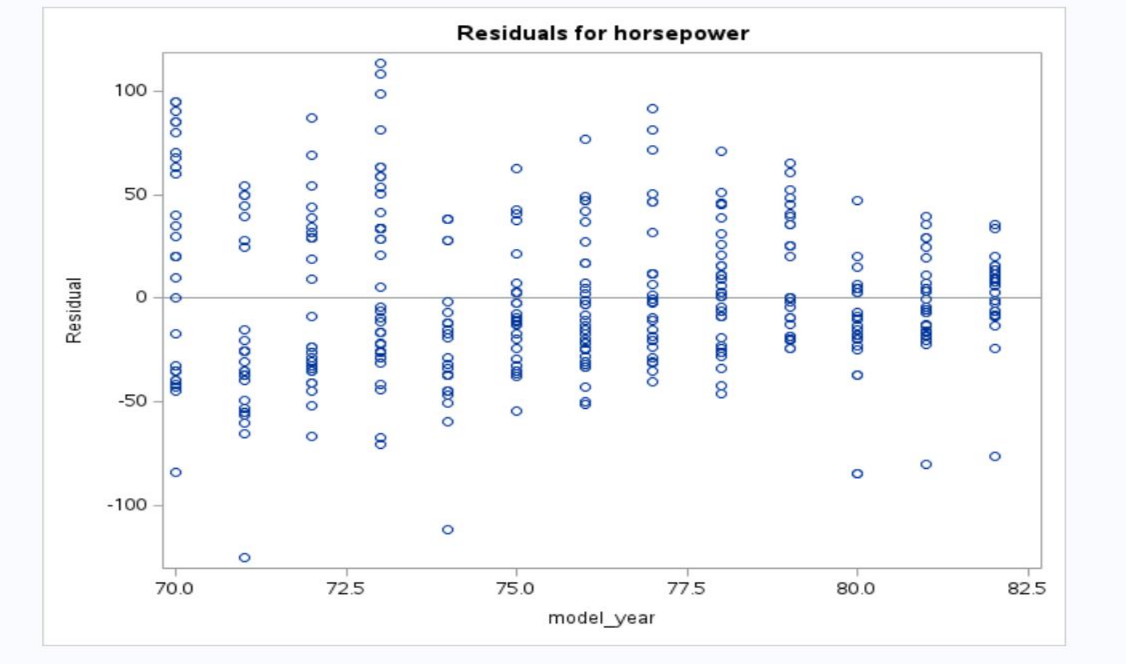
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	443.60600	37.95630	11.69	<.0001
model_year	1	-4.48245	0.49877	-8.99	<.0001

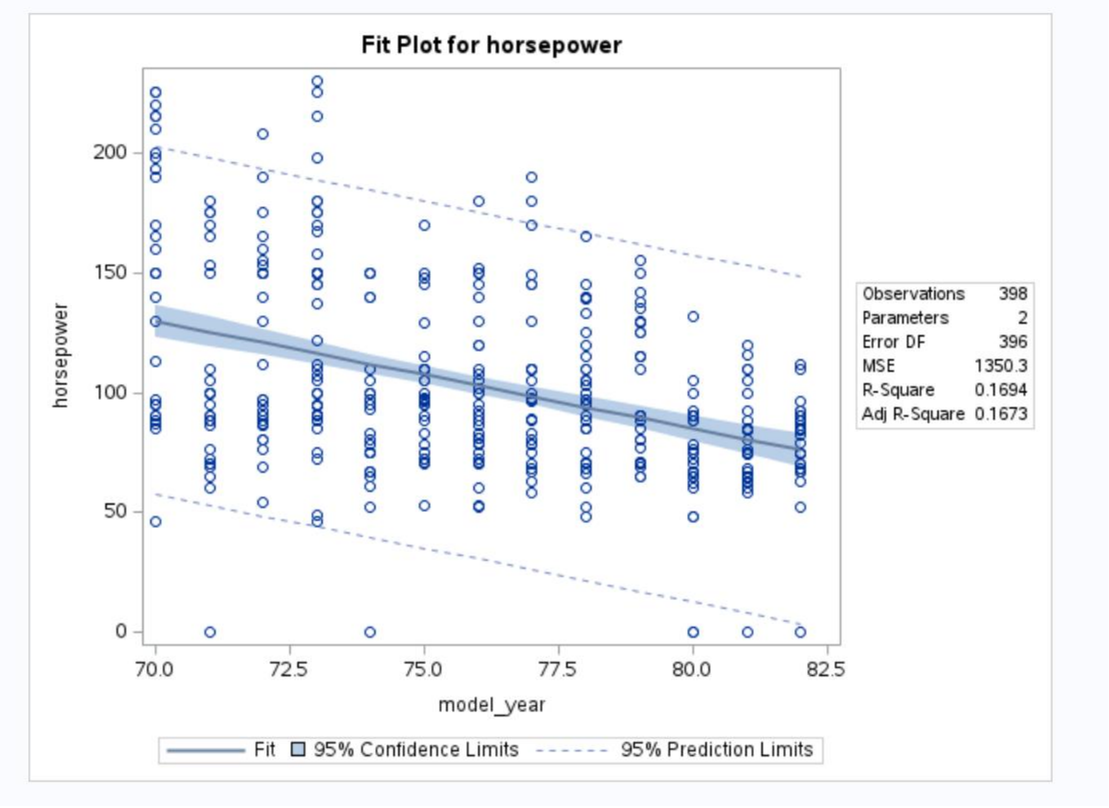


The REG Procedure
Model: MODEL1
Dependent Variable: horsepower

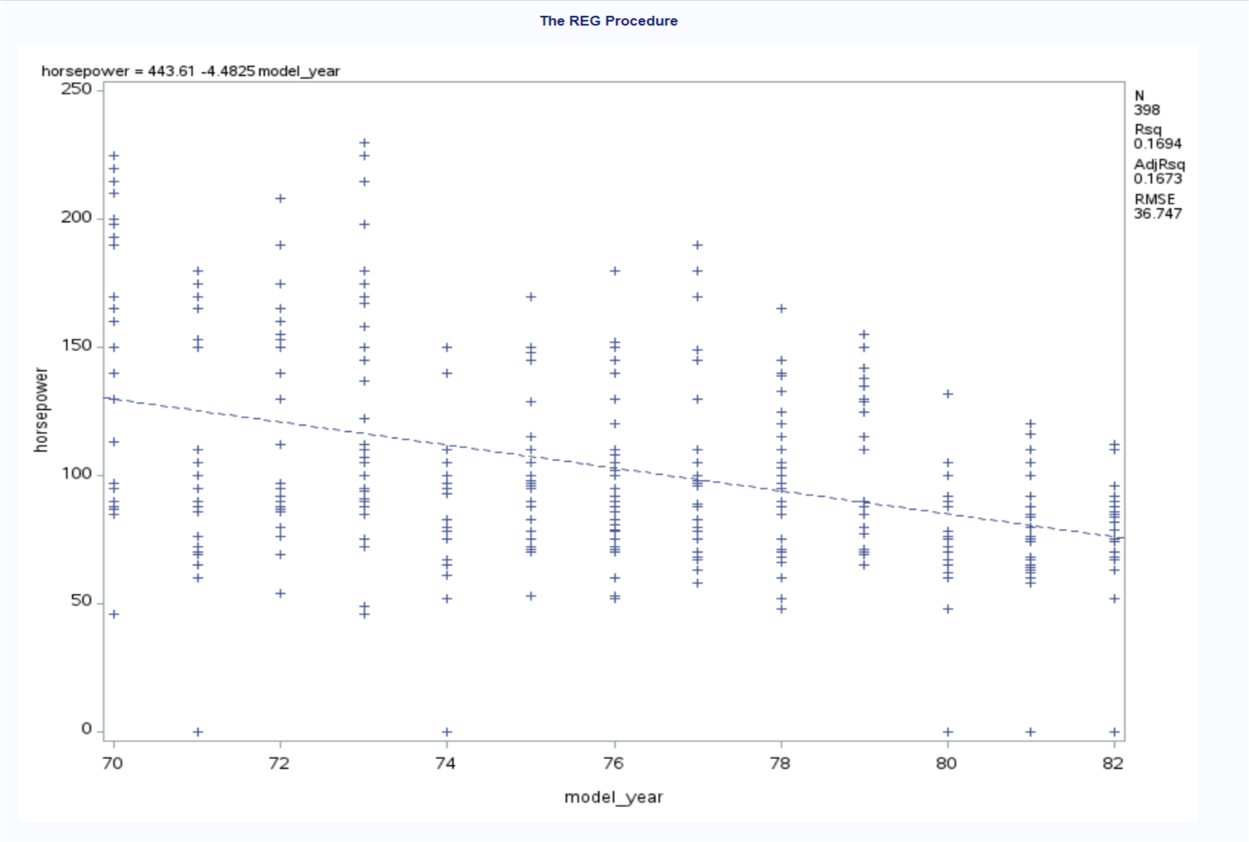
Fit Diagnostics for horsepower

Observations	398
Parameters	2
Error DF	396
MSE	1350.3
R-Square	0.1694
Adj R-Square	0.1673





Observations	398
Parameters	2
Error DF	396
MSE	1350.3
R-Square	0.1694
Adj R-Square	0.1673



The REG Procedure

Answer to the Business Question 2:

The regression analysis suggests that there is a statistically significant relationship between model year and horsepower, with newer models having less horsepower, on average. This could indicate that over the years, there has been a trend toward producing cars with less horsepower.

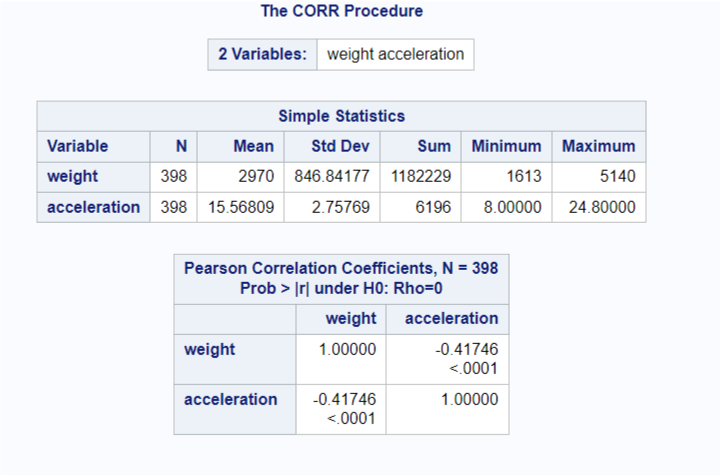
The statistical analysis shows that the model year is a significant predictor of horsepower with an F-value of 80.77, which is highly significant p < .0001. This means that newer model years are associated with lower horsepower, on average.

The R-squared value of 0.1694 suggests that approximately 16.94% of the variation in horsepower is explained by the model year. However, the adjusted R-squared value of 0.1673, which adjusts for the number of predictors in the model, provides a more accurate measure of the relationship's strength. The adjusted R-squared is slightly less than the R-squared value. The regression equation derived from the analysis is horsepower = 443.6060 - 4.48245 * model_year.

Business Question 3

Is there a correlation between the weight of a car and its acceleration?

```
PROC CORR DATA=cleaned_data;
VAR weight acceleration;
RUN;
```



The CORR Procedure

2 Variables:	weight acceleration

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
weight	398	2970	846.84177	1182229	1613	5140
acceleration	398	15.56809	2.75769	6196	8.00000	24.80000

Pearson Correlation Coefficients, N = 398
Prob > |r| under H0: Rho=0

	weight	acceleration
weight	1.00000	-0.41746 <.0001
acceleration	-0.41746 <.0001	1.00000

Answer to the Business Question 3:

Yes, there is a statistically significant correlation between the weight of a car and its acceleration. The negative correlation coefficient indicates that as the weight of a car increases, its acceleration decreases. This is consistent with physical principles, as heavier cars typically require more force to accelerate and generally have slower acceleration rates compared to lighter cars.

Given the correlation coefficient of approximately -0.42, the relationship is moderate, suggesting that while weight is an important factor for acceleration, it is not the sole determinant, and other factors also play a role in a car's acceleration capabilities.