**Assignment 1**

**Name: Jenson Puthenpeedikayin Jacob**

**Student ID: 0794547**

**Section: 2**

**Dataset 1**

**Title: 2022 Fuel Consumption Ratings**

**Source: Kaggle**

**Dataset:** <https://www.kaggle.com/datasets/rinichristy/2022-fuel-consumption-ratings\>

**About the dataset: This Dataset provides model-specific fuel consumption ratings and estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada in 2022.The dataset consist of 16 columns and 14190 rows. Different variables and variable types are available in the dataset which will be explained below:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** **Name** | **Description** | **R Data type** | **Variable types** |
| **Model Year** | **Manufactured year of the vehicle** | **Numerical** | **Ordinal categorical** |
| **Make** | **Vehicle manufacturer name** | **Character** | **Nominal Categorical** |
| **Model** | **Vehicle Model type: 4WD/4X4 = Four-wheel drive; AWD = All-wheel drive; FFV = Flexible-fuel vehicle; SWB = Short** | **Character** | **Nominal Categorical** |
| **Vehicle Class** | **Different class of different branded vehicles like SUV, Mid-Size, Two-seater** | **Character** | **Nominal Categorical** |
| **Engine Size** | **Size of the engine in liters** | **Numerical** | **Continuous** |
| **Cylinders** | **Number of cylinders** | **Numerical** | **Ordinal categorical** |
| **Transmission** | **Transmission: A = automatic; AM = automated manual; AS = automatic with select shift; AV = continuously** | **Character** | **Nominal Categorical** |
| **Fuel type** | **Fuel type: Z – Petrol, D- Diesel, E- Electric.** | **Character** | **Nominal Categorical** |
| **Fuel Consumption (City (L/100 km)** | **City fuel consumption ratings are shown in liters per 100 kilometers (L/100 km)** | **Numerical** | **Continuous** |
| **Fuel Consumption (Hwy (L/100 km))** | **Highway fuel consumption ratings are shown in liters per 100 kilometers (L/100 km)** | **Numerical** | **Continuous** |
| **Fuel Consumption (Comb (L/100 km))** | **The combined rating (55% city, 45% highway) is shown in L/100 km** | **Numerical** | **Continuous** |
| **Fuel Consumption (Comb (mpg))** | **The combined rating (55% city, 45% highway) is shown in miles per imperial gallon (mpg)** | **Numerical** | **Ordinal Categorical** |
| **CO2 Emissions(g/km)** | **The tailpipe emissions of carbon dioxide (in grams per kilometer) for combined city and highway driving** | **Numerical** | **Ordinal Categorical** |
| **CO2 Rating** | **The tailpipe emissions of carbon dioxide rated on a scale from 1 (worst) to 10 (best)** | **Numerical** | **Ordinal Categorical** |
| **Smog Rating** | **The tailpipe emissions of smog-forming pollutants rated on a scale from 1 (worst) to 10 (best)** | **Numerical** | **Ordinal Categorical** |

**Dataset 2**

**Title: Best Selling Mobile Phones**

**Source: Kaggle**

**Dataset:** [**https://www.kaggle.com/datasets/muhammedtausif/best-selling-mobile-phones**](https://www.kaggle.com/datasets/muhammedtausif/best-selling-mobile-phones)

**About the dataset: The dataset contains the data related to best selling smartphones till 2021. This contains of 6 columns and 666 rows.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** **Name** | **Description** | **R Data type** | **Variable types** |
| **Manufacturer** | **Manufacturer of the phone** | **Character** | **Nominal Categorical** |
| **Model** | **Model of the phone** | **Character** | **Nominal Categorical** |
| **Form** | **Type of smartphone. For example, touchscreen or bar phone** | **Character** | **Nominal Categorical** |
| **Smartphone** | **Is the phone smartphone or not?** | **Character** | **Nominal Categorical** |
| **Year** | **Year of launch** | **Numerical** | **Ordinal Categorical** |
| **Units\_sold\_m** | **Total units sold in millions** | **Numerical** | **Ordinal Categorical** |

**Dataset 3**

# Title: Top YouTube Channels Data

# Source: Kaggle

# Dataset: <https://www.kaggle.com/datasets/surajjha101/top-youtube-channels-data>

**About the dataset:** **This data contains the 7 attributes about the top YouTube channels as per number of subscribers they have and 695 rows mentioning about different channels specifications. These attributes with their proper description are as follows:**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Description | R Data type | Variable types |
| Rank | **Rank of the channel as per number of subscribers they have** | Numerical | Ordinal Categorical |
| Youtuber | **Channel Official Name** | Character | Nominal Categorical |
| Subscribers | **Number of subscribers channel have** | **Numerical** | Ordinal Categorical |
| Video Views | **Number for which all videos have been watched collectively** | Numerical | Ordinal Categorical |
| Video count | **Number of videos channel has uploaded so far** | Numerical | Ordinal Categorical |
| Category | **Category (genre) of the channel** | Character | Nominal Categorical |
| Started | **Year when the channel was started** | Numerical | Ordinal Categorical |

**Exercise 2**

**Source:** <https://resagratia.com/2020/07/the-differences-between-good-data-visualization-and-bad-data-visualization-part-1/>

Chart, histogram

Description automatically generated

**Reasons to state the above visualization as a good one:**

1. All the data are properly labelled and displayed in a proper way so that a naïve user can also understand without prior details.
2. Ambiguity and overlay of the graph are not seen.
3. Color combination, proper sorting order, data references are mentioned properly.

**Exercise 3**

**Source**: <https://www.livestories.com/blog/five-ways-to-fail-data-visualization>

A picture containing text, writing implement, stationary, pencil

Description automatically generated

**Reasons to state the above visualization as bad one:**

1. The graph involves too many variables which makes it difficult for the viewer to comprehend it immediately.
2. The 3D design has resulted in undue complexity wherein some bars are simply seemed to be hidden behind those in the front.
3. No labels are found on the bars which makes it difficult to understand.