Trimester II Assignment

Course: MDIMS4DVA Data Visual Analytics (18733)

Portfolio: Data Visual Analytics

Submitted by:
Josiah Olukayode
Student ID No.: 100067107
Faculty of SESS
School of Engineering, Technology and Design

Module Leader:
Dr. Leishi Zhang

Declaration:
I hereby affirm that I have mentioned all my sources and no part of my assignment uses any unacknowledged material.

Table of Contents

1.0 Executive Summary

Today's business world and environment is characterized by increasing volume of data and information generation. In all facets of activities (political, social, media, health care, sales, etc.) big data is being generated on a daily basis and at rapid rate. The need for the application of data visual analytics in solving business problems by creating competitive edge through information processing cannot be overemphasized in this present dispensation of competitive business environment. Before now, the ability of businesses to store, retrieve, process and utilize data to drive business profitably seemed to be a nightmare and difficult task when considering storing and retrieving data for analytical purposes. However, in the last decades, with the emergence of big data generation, the improvement in the data storage via development in the technological sector, and improvement in data visual analytics processes using state of the art tools have made this relatively easy thereby eliminating the problem of storing, retrieving and analyzing data. Data remains valueless if not retrieved, prepared and analyzed to derive actionable insights that can be applied to drive businesses profitably. Data visual analytics is the means through which this can be achieved in order to meet with the competitive demand in the different industries. Visual analytics must be done and applied correctly in order for it to serve its intended purposes. Keim et. al. 2020.

This report aims to apply a blend of existing analytical tools such as Python programming language, Microsoft Excel and Knime Analytics to analyze some selected datasets and making information transparent to discover hidden insights for the user for decision making. The choices and types of dataset includes, but not limited to Tabular Data, Network Data and Geographical Data. Conscious efforts were applied to transform some numerical data to Nominal and Ordinal data. This affords us the opportunity of designing workflow for analyzing, visualizing and reporting of the findings from the dataset.

2. 0 Brief History Of Data Visual Analytics

Graphical presentation of data cannot be said to be a relatively modern development in statistics. It moved from confirmatory data analysis to exploratory data analysis. John W. Tukey (1977) in his book titled “Exploratory Data Analysis” stated that the perceived limitations of the existing traditional statistical techniques (i.e. automatic analysis techniques), which developed separately from data visualization and interactive techniques necessitated one of the moves to modern day visual analytics research. This was as a result of the limited scope of the automatic analysis techniques. According to P.C Wong (2004) and J. Thomas (2005), the term visual analytics were mentioned twice in IEEE (Institute of Electrical and Electronics Engineers) Computer Graphics and Applications and research and development agenda titled “Illuminating the Path”. The more recent usage of the term in a wider context was in a new multidisciplinary field which combines several areas of research that includes visualization, data management, data analysis, geo-spatial and temporal data processing, spatial decisions support and statistics. Data visual analytics became more easy with technological advancement resulting from the discovery of the need for potential integration of knowledge discovery and data mining process by a research community which devoted their efforts to data visualization. Consequently, it made available an integration of softwares with modern day devices that have interactive data visualization functionalities and capabilities. This allowed for an easy visual data exploration, knowledge discovery, data mining process, and knowledge transfer. Keim et. al. 2020

2.1 Visual Analytics.

The multi-disciplinary nature (which involves multiple processes and wide range of application areas) of visual analytics may serve as a determining factor in defining this term. Before now, P.C Wong (2004) defined it as "the science of analytical reasoning facilitated by interactive human-machine interfaces". According to Keim et. al. (2020), a more recent definition informed by a current practice gave a more specific definition as “Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets". From the stand point of the goal of visual analytics, a more amplification of this definition is permitted to state that visual analytics is the formation of tools and techniques which enables professionals to harmonize information and derive actionable insights from big, changing, unclear, and often conflicting datasets. It allows for detection of the expected and discovery of the unexpected. It provides timely, defensible, and understandable assessments of the findings, and effective communication of these findings for benefitable action. Keim et. al. (2020)

2.2 Visual vs Non Visual Data Analysis (Activity 2.1)

There are two ways through which data can be analyzed, interpreted and communicate findings in an effective way for decision making. This section of the portfolio discusses the no-visual way of analyzing data.

The non-visual group was set out to find non-trivial and interesting findings from the provided car dataset. This section of this portfolio utilizes a combination of existing tools for data analysis which are excel and python programming language as the analytic tool.

Excel was used to do some analysis on the car dataset. The functionalities that was used in the excel domain was power query editor and pivot table while integrated development environment used for python programming language was Jupyter notebook.

2.2.1 Car Data Dataset Description (Using Python)

```
[4]: import pandas as pd

[5]: df = pd.read_csv("data_cars.csv")

[6]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 428 entries, 0 to 427
Data columns (total 20 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Vehicle Name     428 non-null    object
 1   Sedan            428 non-null    int64
 2   Sports Car       428 non-null    int64
 3   SUV              428 non-null    int64
 4   Wagon            428 non-null    int64
 5   Minivan          428 non-null    int64
 6   Pickup           428 non-null    int64
 7   AWD              428 non-null    int64
 8   RWD              428 non-null    int64
 9   Retail Price     428 non-null    int64
 10  Dealer Cost      428 non-null    int64
 11  Engine Size (l)  428 non-null    float64
 12  Cyl              428 non-null    int64
 13  HP               428 non-null    int64
 14  City MPG         428 non-null    object
 15  Hwy MPG          428 non-null    object
 16  Weight           428 non-null    object
 17  Wheel Base       428 non-null    object
 18  Len              428 non-null    object
 19  Width            428 non-null    object
dtypes: float64(1), int64(12), object(7)
memory usage: 67.0+ KB
```

Figure 1: Car dataset description

2.2.2 Car dataset Statistical Measures

Statistical measures are calculated using the python df.describe() returned the following values for each of the dataset attributes. For each of the attribute, the python data frame describe function provided some statistical measure of the car dataset.

[4]:

	Sedan	Sports Car	SUV	Wagon	Minivan	Pickup
count	428.000000	428.000000	428.000000	428.000000	428.000000	428.000000
mean	0.572430	0.114486	0.140187	0.070093	0.046729	0.056075
std	0.495305	0.318773	0.347587	0.255603	0.211305	0.230335
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

	AWD	RWD	Retail Price	Dealer Cost	Engine Size (l)
count	428.000000	428.000000	428.000000	428.000000	428.000000
mean	0.214953	0.257009	32774.855140	30014.700935	3.196729
std	0.411271	0.437496	19431.716674	17642.117750	1.108595
min	0.000000	0.000000	10280.000000	9875.000000	1.300000
25%	0.000000	0.000000	20334.250000	18866.000000	2.375000
50%	0.000000	0.000000	27635.000000	25294.500000	3.000000
75%	0.000000	1.000000	39205.000000	35710.250000	3.900000
max	1.000000	1.000000	192465.000000	173560.000000	8.300000

	Cyl	HP
count	428.000000	428.000000
mean	5.775701	215.885514
std	1.622779	71.836032
min	-1.000000	73.000000
25%	4.000000	165.000000
50%	6.000000	210.000000
75%	6.000000	255.000000
max	12.000000	500.000000

Figure 2: Statistical Measures for the Car Dataset

The interpretation of the statistical measures were:

1. Count : There were a total count of 428 data points (rows) for each attribute.
2. Mean (Average) for each attribute
3. Standard Deviation (std) for each attribute
4. Minimum (min) values for each attribute
5. 25th percentile (25%) for each attribute
6. 50th percentile (50%) for each attribute
7. 75th percentile for each attribute
8. Maximum (max) value for each attribute

2.2.3 Non-Visual Analysis of Car Dataset Attributes

The car dataset was subjected to an initial correlation analysis. Listed below are the attributes which were further subjected to further analysis:

A. Retail Price
B. Dealer Cost
C. HP (Horse Power)
D. Hwy MPG (Highway Miles per gallon)
E. City MPG (City Miles per gallon)
F. Cylinder
G. Engine Size

```
import numpy as np

df.corr()
```

	Sedan	Sports Car	SUV	Wagon	Minivan	Pickup	AWD	RWD	Retail Price	Dealer Cost	Engine Size (l)	Cyl	HP
Sedan	1.000000	-0.416041	-0.467207	-0.317670	-0.256178	-0.282015	-0.318038	-0.096914	-0.176489	-0.168657	-0.301108	-0.174956	-0.254785
Sports Car	-0.416041	1.000000	-0.145188	-0.098718	-0.079609	-0.087638	-0.098833	0.393055	0.381856	0.376644	0.079924	0.058811	0.342155
SUV	-0.467207	-0.145188	1.000000	-0.110859	-0.089400	-0.098416	0.411247	-0.237484	0.041928	0.036907	0.263747	0.197042	0.112163
Wagon	-0.317670	-0.098718	-0.110859	1.000000	-0.060786	-0.066917	0.056840	-0.014875	-0.055653	-0.052491	-0.105805	-0.080575	-0.083742
Minivan	-0.256178	-0.079609	-0.089400	-0.060786	1.000000	-0.053963	-0.035008	-0.104884	-0.056789	-0.058540	0.067637	0.016979	-0.035286
Pickup	-0.282015	-0.087638	-0.098416	-0.066917	-0.053963	1.000000	0.169126	0.135531	-0.098371	-0.102325	0.194238	0.071321	0.030395
AWD	-0.318038	-0.098833	0.411247	0.056840	-0.035008	0.169126	1.000000	-0.307756	0.099985	0.094665	0.203412	0.142591	0.140110
RWD	-0.096914	0.393055	-0.237484	-0.014875	-0.104884	0.135531	-0.307756	1.000000	0.403593	0.403410	0.264899	0.299100	0.382689
Retail Price	-0.176489	0.381856	0.041928	-0.055653	-0.056789	-0.098371	0.099985	0.403593	1.000000	0.999132	0.571753	0.628757	0.826945
Dealer Cost	-0.168657	0.376644	0.036907	-0.052491	-0.058540	-0.102325	0.094665	0.403410	0.999132	1.000000	0.564498	0.624204	0.823746
Engine Size (l)	-0.301108	0.079924	0.263747	-0.105805	0.067637	0.194238	0.203412	0.264899	0.571753	0.564498	1.000000	0.897564	0.787435
Cyl	-0.174956	0.058811	0.197042	-0.080575	0.016979	0.071321	0.142591	0.299100	0.628757	0.624204	0.897564	1.000000	0.775799
HP	-0.254785	0.342155	0.112163	-0.083742	-0.035286	0.030395	0.140110	0.382689	0.826945	0.823746	0.787435	0.775799	1.000000

Figure 3: General Correlation Analysis on all Car Dataset Attributes

Figure 3 shows the initial correlation analysis on all the car dataset attributes. Based on the outcome of the analysis, seven (7) were discovered to have statistically significant relationship based on their correlation coefficient (r).

These attributes were analyzed in no any order of importance (order of importance depends on the audience's preference).

The categories of vehicles analyzed by the listed attributes were:

A. Minivan
B. Pickup
C. Sedan
D. Sports Car
E. SUV
F. Wagon

In the analysis of the car dataset using excel, the retail price was used to filter (from Highest to Lowest) while in terms of fuel consumption, the City MPG was used to filter (from lowest to highest). Below are the summary of analysis and non-trivial findings from the dataset:

Analysis of the Minivan Category

Retail Price and Dealer Cost Analysis (Minivan Category)		
Stock Status	Available	
Vehicle Name/Type	**Retail Price**	**Dealer Cost**
Chrysler Town and Country Limited/Minivan	£38,380	£35,063
Mercury Monterey Luxury/Minivan	£33,995	£30,846
Nissan Quest SE/Minivan	£32,780	£30,019
Dodge Grand Caravan SXT/Minivan	£32,660	£29,812
Pontiac Montana EWB/Minivan	£31,370	£28,454
Toyota Sienna XLE Limited/Minivan	£28,800	£25,690
Oldsmobile Silhouette GL/Minivan	£28,790	£26,120
Mazda MPV ES/Minivan	£28,750	£26,600
Chrysler Town and Country LX/Minivan	£27,490	£25,371
Honda Odyssey EX/Minivan	£27,450	£24,744
Chevrolet Venture LS/Minivan	£27,020	£24,518
Ford Freestar SE/Minivan	£26,930	£24,498
Chevrolet Astro/Minivan	£26,395	£23,954
GMC Safari SLE/Minivan	£25,640	£23,215
Honda Odyssey LX/Minivan	£24,950	£22,498
Nissan Quest S/Minivan	£24,780	£22,958
Pontiac Montana/Minivan	£23,845	£21,644
Toyota Sienna CE/Minivan	£23,495	£21,198
Dodge Caravan SE/Minivan	£21,795	£20,508
Kia Sedona LX/Minivan	£20,615	£19,400

Engine Size and Mile Per Gallon Analysis (Minivan Category)					
Stock Status	Available				
Vehicle Name/Type	**City MPG**	**Hwy MPG**	**Engine Size (l)**	**Cyl**	**HP**
Chevrolet Astro/Minivan	14	17	4.3	6	190
GMC Safari SLE/Minivan	16	20	4.3	6	190
Kia Sedona LX/Minivan	16	22	3.5	6	195
Mercury Monterey Luxury/Minivan	16	23	4.2	6	201
Ford Freestar SE/Minivan	17	23	3.9	6	193
Honda Odyssey LX/Minivan	18	25	3.5	6	240
Dodge Grand Caravan SXT/Minivan	18	25	3.8	6	215
Mazda MPV ES/Minivan	18	25	3	6	200
Pontiac Montana EWB/Minivan	18	24	3.4	6	185
Chrysler Town and Country Limited/Minivan	18	25	3.8	6	215
Honda Odyssey EX/Minivan	18	25	3.5	6	240
Nissan Quest SE/Minivan	18	25	3.5	6	240
Nissan Quest S/Minivan	19	26	3.5	6	240
Chevrolet Venture LS/Minivan	19	26	3.4	6	185
Chrysler Town and Country LX/Minivan	19	26	3.3	6	180
Toyota Sienna XLE Limited/Minivan	19	27	3.3	6	230
Pontiac Montana/Minivan	19	26	3.4	6	185
Oldsmobile Silhouette GL/Minivan	19	26	3.4	6	185
Toyota Sienna CE/Minivan	19	27	3.3	6	230
Dodge Caravan SE/Minivan	20	26	2.4	4	150

Figure 4: Analysis for Minivan Category.

From figure 4, which indicates the analysis of the Minivan Category, it was discovered that for:

a. **Retail Price and Dealer Cost**: A 6 cylinders Chrysler Town and Country Limited/Minivan ranked the topmost with the highest retail price of £38, 380 and Dealer Cost of £35, 063 while the Kia Sedona LX/ Minivan ranked the lowest on retail price of £20, 615 and Dealer Cost of £19,400 respectively with 6 cylinders and an engine size of 3.3 liters.
b. **Engine Size, Miles per gallon and Cylinder size**: It was discovered that Chevrolet Astro/Minivan ranked the topmost with an engine size of 4.3 liters however, considering city MPG, it ranked the topmost in that it has the lowest city miles per gallon of 14 while Dodge Caravan SE/Minivan with the lowest engine capacity of 2.4 liters in the category ranked the lowest in the City MPG and second lowest in the Hwy MPG consideration (i.e. 20 liters for Hwy MPG and 26 liters City MPG).

Analysis of the Pickup Category

Retail Price and Dealer Cost Analysis (Pickup Category)		
Stock Status	Available	
Vehicle Name/Type	**Retail Price**	**Dealer Cost**
Cadillac Escalade EXT/Pickup	£52,975	£48,541
Chevrolet SSR/Pickup	£41,995	£39,306
Chevrolet Silverado SS/Pickup	£40,340	£35,399
Chevrolet Avalanche 1500/Pickup	£36,100	£31,689
Ford F-150 Supercab Lariat/Pickup	£33,540	£29,405
GMC Sierra HD 2500/Pickup	£29,322	£25,759
Nissan Titan King Cab XE/Pickup	£26,650	£24,926
Toyota Tundra Access Cab V6 SR5/Pickup	£25,935	£23,520
GMC Sierra Extended Cab 1500/Pickup	£25,717	£22,604
GMC Sonoma Crew Cab/Pickup	£25,395	£23,043
Subaru Baja/Pickup	£24,520	£22,304
Mazda B4000 SE Cab Plus/Pickup	£22,350	£20,482
Ford F-150 Regular Cab XL/Pickup	£22,010	£19,490
Chevrolet Silverado 1500 Regular Cab/Pickup	£20,310	£18,480
Dodge Dakota Club Cab/Pickup	£20,300	£18,670
Dodge Ram 1500 Regular Cab ST/Pickup	£20,215	£18,076
Nissan Frontier King Cab XE V6/Pickup	£19,479	£18,253
Chevrolet Colorado Z85/Pickup	£18,760	£17,070
Dodge Dakota Regular Cab/Pickup	£17,630	£16,264

Engine Size and Mile Per Gallon Analysis (Pickup Category)					
Stock Status	Available				
Vehicle Name/Type	**Hwy MPG**	**City MPG**	**Engine Size (l)**	**Cyl**	**HP**
GMC Sierra HD 2500/Pickup			6	8	300
Cadillac Escalade EXT/Pickup	17	13	6	8	345
Chevrolet Silverado SS/Pickup	17	13	6	8	300
Toyota Tundra Access Cab V6 SR5/Pickup	17	14	3.4	6	190
Ford F-150 Supercab Lariat/Pickup	18	14	5.4	8	300
Chevrolet Avalanche 1500/Pickup	18	14	5.3	8	295
Nissan Titan King Cab XE/Pickup	18	14	5.6	8	305
Ford F-150 Regular Cab XL/Pickup	19	15	4.6	8	231
GMC Sonoma Crew Cab/Pickup	19	15	4.3	6	190
Mazda B4000 SE Cab Plus/Pickup	19	15	4	6	207
Chevrolet SSR/Pickup	19	16	5.3	8	300
GMC Sierra Extended Cab 1500/Pickup	20	17	4.8	8	285
Nissan Frontier King Cab XE V6/Pickup	20	17	3.3	6	180
Dodge Ram 1500 Regular Cab ST/Pickup	21	16	3.7	6	215
Chevrolet Silverado 1500 Regular Cab/Pickup	21	15	4.3	6	200
Dodge Dakota Club Cab/Pickup	22	16	3.7	6	210
Dodge Dakota Regular Cab/Pickup	22	16	3.7	6	210
Chevrolet Colorado Z85/Pickup	23	18	2.8	4	175
Subaru Baja/Pickup	28	21	2.5	4	165

Figure 5: Analysis for Pickup Category.

From figure 5, which indicates the analysis of the Pickup Category, it was discovered that for:

a. **Retail Price and Dealer Cost**: Cadillac Escalade EXT/Pickup ranked the topmost with the highest retail price of £52, 975 and Dealer Cost of £48, 541, while Dodge Dakota Regular Cab/Pickup ranked the lowest on retail price of £17, 630 and Dealer Cost of £16,4264 respectively.
b. **Engine Size, Miles per gallon and Cylinder size**: It was discovered that GMC Sierra HD 2500/Pickup that have 8 cylinders ranked the topmost with an engine size of 6 liters however, the Hwy MPG and City MPG were not reported. A 4 cylinder Subaru Baja/Pickup with Engine Size of 2.5 liters ranked lowest in terms of miles per gallon in that it has the highest highway miles per gallon of 28 and city miles per gallon 21.

Analysis for Sports Car Category

Retail Price and Dealer Cost Analysis (Sports Car Category)		
Stock Status	Available	
Vehicle Name/Type	**Retail Price**	**Dealer Cost**
Porsche 911 GT2 2dr/Sports Car	£192,465	£173,560
Mercedes-Benz SL600 convertible 2dr/Sports Car	£126,670	£117,854
Mercedes-Benz SL55 AMG 2dr/Sports Car	£121,770	£113,388
Mercedes-Benz SL500 convertible 2dr/Sports Car	£90,520	£84,325
Acura NSX coupe 2dr manual S/Sports Car	£89,765	£79,978
Jaguar XKR convertible 2dr/Sports Car	£86,995	£79,226
Audi RS 6 4dr/Sports Car	£84,600	£76,417
Porsche 911 Carrera 4S coupe 2dr (convert)/Sports Car	£84,165	£72,206
Jaguar XKR coupe 2dr/Sports Car	£81,995	£74,676
Dodge Viper SRT-10 convertible 2dr/Sports Car	£81,795	£74,451
Porsche 911 Carrera convertible 2dr (coupe)/Sports Car	£79,165	£69,229
Porsche 911 Targa coupe 2dr/Sports Car	£76,765	£67,128
Cadillac XLR convertible 2dr/Sports Car	£76,200	£70,546
Jaguar XK8 convertible 2dr/Sports Car	£74,995	£68,306
Jaguar XK8 coupe 2dr/Sports Car	£69,995	£63,756
Lexus SC 430 convertible 2dr/Sports Car	£63,200	£55,063
BMW M3 convertible 2dr/Sports Car	£56,595	£51,815
Mercedes-Benz SLK32 AMG 2dr/Sports Car	£56,170	£52,289
Porsche Boxster S convertible 2dr/Sports Car	£52,365	£45,766
Chevrolet Corvette convertible 2dr/Sports Car	£51,535	£45,193
BMW M3 coupe 2dr/Sports Car	£48,195	£44,170
Chevrolet Corvette 2dr/Sports Car	£44,535	£39,068
Porsche Boxster convertible 2dr/Sports Car	£43,365	£37,886
BMW Z4 convertible 3.0i 2dr/Sports Car	£41,045	£37,575
Audi TT 3.2 coupe 2dr (convertible)/Sports Car	£40,590	£36,739
Mercedes-Benz SLK230 convertible 2dr/Sports Car	£40,320	£37,548
Ford Thunderbird Deluxe convert w/hardtop 2dr/Sports Car	£37,530	£34,483
Audi TT 1.8 Quattro 2dr (convertible)/Sports Car	£37,390	£33,891
Audi TT 1.8 convertible 2dr (coupe)/Sports Car	£35,940	£32,512
Chrysler Crossfire 2dr/Sports Car	£34,495	£32,033
Nissan 350Z Enthusiast convertible 2dr/Sports Car	£34,390	£31,845
BMW Z4 convertible 2.5i 2dr/Sports Car	£33,895	£31,065
Pontiac GTO 2dr/Sports Car	£33,500	£30,710
Honda S2000 convertible 2dr/Sports Car	£33,260	£29,965
Subaru Impreza WRX STi 4dr/Sports Car	£31,545	£29,130
Mitsubishi Lancer Evolution 4dr/Sports Car	£29,562	£27,466
Ford Mustang GT Premium convertible 2dr/Sports Car	£29,380	£26,875
Mazda RX-8 4dr manual/Sports Car	£27,200	£25,179
Mitsubishi Eclipse Spyder GT convertible 2dr/Sports Car	£26,992	£25,218
Nissan 350Z coupe 2dr/Sports Car	£26,910	£25,203
Mazda RX-8 4dr automatic/Sports Car	£25,700	£23,794
Mazda MX-5 Miata LS convertible 2dr/Sports Car	£25,193	£23,285
Toyota MR2 Spyder convertible 2dr/Sports Car	£25,130	£22,787
Mitsubishi Eclipse GTS 2dr/Sports Car	£25,092	£23,456
Subaru Impreza WRX 4dr/Sports Car	£25,045	£23,022
Toyota Celica GT-S 2dr/Sports Car	£22,570	£20,363
Mazda MX-5 Miata convertible 2dr/Sports Car	£22,388	£20,701
Hyundai Tiburon GT V6 2dr/Sports Car	£18,739	£17,101
Ford Mustang 2dr (convertible)/Sports Car	£18,345	£16,943

Engine Size and Mile Per Gallon Analysis (Sports Car Category)					
Stock Status	Available				
Vehicle Name/Type	**City MPG**	**Hwy MPG**	**Engine Size (l)**	**Cyl**	**HP**
Dodge Viper SRT-10 convertible 2dr/Sports Car			8.3	10	500
Pontiac GTO 2dr/Sports Car			5.7	8	340
Mercedes-Benz SL600 convertible 2dr/Sports Car	13	19	5.5	12	493
Mercedes-Benz SL55 AMG 2dr/Sports Car	14	21	5.5	8	493
Audi RS 6 4dr/Sports Car	15	22	4.2	8	450
Jaguar XKR coupe 2dr/Sports Car	16	23	4.2	8	390
Jaguar XKR convertible 2dr/Sports Car	16	23	4.2	8	390
BMW M3 coupe 2dr/Sports Car	16	24	3.2	6	333
BMW M3 convertible 2dr/Sports Car	16	23	3.2	6	333
Mercedes-Benz SL500 convertible 2dr/Sports Car	16	23	5	8	302
Acura NSX coupe 2dr manual S/Sports Car	17	24	3.2	6	290
Ford Mustang GT Premium convertible 2dr/Sports Car	17	25	4.6	8	260
Porsche 911 GT2 2dr/Sports Car	17	24	3.6	6	477
Mercedes-Benz SLK32 AMG 2dr/Sports Car	17	22	3.2	6	349
Chrysler Crossfire 2dr/Sports Car	17	25	3.2	6	215
Ford Thunderbird Deluxe convert w/hardtop 2dr/Sports Car	17	24	3.9	8	280
Cadillac XLR convertible 2dr/Sports Car	17	25	4.6	8	320
Porsche 911 Carrera 4S coupe 2dr (convert)/Sports Car	17	24	3.6	6	315
Mazda RX-8 4dr automatic/Sports Car	18	25	1.3	-1	197
Mazda RX-8 4dr manual/Sports Car	18	24	1.3	-1	238
Jaguar XK8 convertible 2dr/Sports Car	18	26	4.2	8	294
Jaguar XK8 coupe 2dr/Sports Car	18	26	4.2	8	294
Mitsubishi Lancer Evolution 4dr/Sports Car	18	26	2	4	271
Subaru Impreza WRX STi 4dr/Sports Car	18	24	2.5	4	300
Chevrolet Corvette 2dr/Sports Car	18	25	5.7	8	350
Porsche Boxster S convertible 2dr/Sports Car	18	26	3.2	6	258
Porsche 911 Carrera convertible 2dr (coupe)/Sports Car	18	26	3.6	6	315
Lexus SC 430 convertible 2dr/Sports Car	18	23	4.3	8	300
Porsche 911 Targa coupe 2dr/Sports Car	18	26	3.6	6	315
Chevrolet Corvette convertible 2dr/Sports Car	18	25	5.7	8	350
Hyundai Tiburon GT V6 2dr/Sports Car	19	26	2.7	6	172
BMW Z4 convertible 2.5i 2dr/Sports Car	20	28	2.5	6	184
Honda S2000 convertible 2dr/Sports Car	20	25	2.2	4	240
Ford Mustang 2dr (convertible)/Sports Car	20	29	3.8	6	193
Subaru Impreza WRX 4dr/Sports Car	20	27	2	4	227
Audi TT 1.8 Quattro 2dr (convertible)/Sports Car	20	28	1.8	4	225
Nissan 350Z Enthusiast convertible 2dr/Sports Car	20	26	3.5	6	287
Porsche Boxster convertible 2dr/Sports Car	20	29	2.7	6	228
Nissan 350Z coupe 2dr/Sports Car	20	26	3.5	6	287
Audi TT 1.8 convertible 2dr (coupe)/Sports Car	20	28	1.8	4	180
Mitsubishi Eclipse GTS 2dr/Sports Car	21	28	3	6	210
Mitsubishi Eclipse Spyder GT convertible 2dr/Sports Car	21	28	3	6	210
BMW Z4 convertible 3.0i 2dr/Sports Car	21	29	3	6	225
Mercedes-Benz SLK230 convertible 2dr/Sports Car	21	29	2.3	4	192
Audi TT 3.2 coupe 2dr (convertible)/Sports Car	21	29	3.2	6	250
Mazda MX-5 Miata convertible 2dr/Sports Car	23	28	1.8	4	142
Mazda MX-5 Miata LS convertible 2dr/Sports Car	23	28	1.8	4	142
Toyota Celica GT-S 2dr/Sports Car	24	33	1.8	4	180
Toyota MR2 Spyder convertible 2dr/Sports Car	26	32	1.8	4	138

Figure 6: Analysis for Sport Car Category.

From figure 6, which indicates the analysis of the Sport Car Category, it was discovered that for:

Retail Price and Dealer Cost: Porsche 911 GT2 2dr/Sports Car ranked the topmost with the highest retail price of £192, 465 and Dealer Cost of £173, 560, while Ford Mustang 2dr (convertible)/Sports Car ranked the lowest on retail price of £18, 345 and Dealer Cost of £16, 943 respectively.

Engine Size, Miles per gallon and Cylinder size: It was discovered that Dodge Mercedes-Benz SL600 convertible 2dr/Sports Car that have 12 cylinders and Engine size of 5.5 ranked the topmost with the lowest City MPG of 13 and Hwy MPG of 19 respectively. A 4 cylinder Toyota MR2 Spyder convertible 2dr/Sports Car with Engine Size of 1.8 liters ranked lowest in terms of miles per gallon in that it has the highest City MPG of 26 and Hwy MPG of 32 respectively.

Analysis for SUV Category

Retail Price and Dealer Cost Analysis (Sports Car Category)		
Stock Status	Available	
Vehicle Name/Type	**Retail Price**	**Dealer Cost**
Mercedes-Benz G500/SUV	£76,870	£71,540
Land Rover Range Rover HSE/SUV	£72,250	£65,807
Lexus LX 470/SUV	£64,800	£56,455
Porsche Cayenne S/SUV	£56,665	£49,865
Toyota Land Cruiser/SUV	£54,765	£47,986
Cadillac Escaladet/SUV	£52,795	£48,377
Lincoln Navigator Luxury/SUV	£52,775	£46,360
BMW X5 4.4i/SUV	£52,195	£47,720
Hummer H2/SUV	£49,995	£45,815
Cadillac SRX V8/SUV	£46,995	£43,523
Mercedes-Benz ML500/SUV	£46,470	£43,268
GMC Yukon XL 2500 SLT/SUV	£46,265	£40,534
Lexus GX 470/SUV	£45,700	£39,838
Lincoln Aviator Ultimate/SUV	£42,915	£39,443
Chevrolet Suburban 1500 LT/SUV	£42,735	£37,422
Ford Excursion 6.8 XLT/SUV	£41,475	£36,494
Chevrolet Tahoe LT/SUV	£41,465	£36,287
Volvo XC90 T6/SUV	£41,250	£38,851
Land Rover Discovery SE/SUV	£39,250	£35,777
Lexus RX 330/SUV	£39,195	£34,576
Buick Rainier/SUV	£37,895	£34,357
BMW X3 3.0i/SUV	£37,000	£33,873
Acura MDX/SUV	£36,945	£33,337
CMC Yukon 1500 SLE/SUV	£35,725	£31,361
Toyota Sequoia SR5/SUV	£35,695	£31,827
Volkswagen Touareg V6/SUV	£35,515	£32,243
Ford Expedition 4.6 XLT/SUV	£34,560	£30,468
Nissan Pathfinder Armada SE/SUV	£33,840	£30,815
Mitsubishi Montero XLS/SUV	£33,112	£30,763
Dodge Durango SLT/SUV	£32,235	£29,472
GMC Envoy XUV SLE/SUV	£31,890	£28,922
Isuzu Ascender S/SUV	£31,849	£29,977
Mitsubishi Endeavor XLS/SUV	£30,492	£28,330
Chevrolet TrailBlazer LT/SUV	£30,295	£27,479
Mercury Mountaineer/SUV	£29,995	£27,317
Ford Explorer XLT V6/SUV	£29,670	£26,983
Toyota Highlander V6/SUV	£27,930	£24,915
Jeep Grand Cherokee Laredo/SUV	£27,905	£25,686
Toyota 4Runner SR5 V6/SUV	£27,710	£24,801
Honda Pilot LX/SUV	£27,560	£24,843
Nissan Pathfinder SE/SUV	£27,339	£25,972
Buick Rendezvous CX/SUV	£26,545	£24,085
Land Rover Freelander SE/SUV	£25,995	£23,969
Jeep Wrangler Sahara convertible 2dr/SUV	£25,520	£23,275
Suzuki XL-7 EX/SUV	£23,699	£22,307
Ford Escape XLS/SUV	£22,515	£20,907
Pontiac Aztekt/SUV	£21,595	£19,810
Hyundai Santa Fe GLS/SUV	£21,589	£20,201
Mazda Tribute DX 2.0/SUV	£21,087	£19,742
Nissan Xterra XE V6/SUV	£20,939	£19,512
Saturn VUE/SUV	£20,585	£19,238
Isuzu Rodeo S/SUV	£20,449	£19,261
Toyota RAV4/SUV	£20,290	£18,553
Chevrolet Tracker/SUV	£20,255	£19,108
Jeep Liberty Sport/SUV	£20,130	£18,973
Honda CR-V LX/SUV	£19,860	£18,419
Kia Sorento LX/SUV	£19,635	£18,630
Mitsubishi Outlander LS/SUV	£18,892	£17,569
Honda Element LX/SUV	£18,690	£17,334

Engine Size and Mile Per Gallon Analysis (Sports Car Category)					
Stock Status	Available				
Vehicle Name/Type	**City MPG**	**Hwy MPG**	**Engine Size (l)**	**Cyl**	**HP**
Ford Excursion 6.8 XLT/SUV			6.8	10	310
Hummer H2/SUV	10	12	6	8	316
Land Rover Discovery SE/SUV	12	16	4.6	8	217
Land Rover Range Rover HSE/SUV	12	16	4.4	8	282
Lincoln Aviator Ultimate/SUV	13	18	4.6	8	302
GMC Yukon XL 2500 SLT/SUV	13	17	6	8	325
Lincoln Navigator Luxury/SUV	13	18	5.4	8	300
Lexus LX 470/SUV	13	17	4.7	8	235
Mercedes-Benz G500/SUV	13	14	5	8	292
Toyota Land Cruiser/SUV	13	17	4.7	8	325
Nissan Pathfinder Armada SE/SUV	13	19	5.6	8	305
Toyota Sequoia SR5/SUV	14	17	4.7	8	240
Chevrolet Suburban 1500 LT/SUV	14	18	5.3	8	295
Cadillac Escaladet/SUV	14	18	5.3	8	295
Porsche Cayenne S/SUV	14	18	4.5	8	340
Mercedes-Benz ML500/SUV	14	17	5	8	288
Chevrolet Tahoe LT/SUV	14	18	5.3	8	295
Lexus GX 470/SUV	15	19	4.7	8	235
Ford Expedition 4.6 XLT/SUV	15	19	4.6	8	232
Buick Rainier/SUV	15	21	4.2	6	275
Ford Explorer XLT V6/SUV	15	20	4	6	210
Volvo XC90 T6/SUV	15	20	2.9	6	268
GMC Envoy XUV SLE/SUV	15	19	4.2	6	275
Isuzu Ascender S/SUV	15	20	4.2	6	275
Dodge Durango SLT/SUV	15	21	4.7	8	230
Mitsubishi Montero XLS/SUV	15	19	3.8	6	215
Volkswagen Touareg V6/SUV	15	20	3.2	6	220
Kia Sorento LX/SUV	16	19	3.5	6	192
Mercury Mountaineer/SUV	16	21	4	6	210
Jeep Grand Cherokee Laredo/SUV	16	21	4	6	195
Jeep Wrangler Sahara convertible 2dr/SUV	16	19	4	6	190
BMW X3 3.0i/SUV	16	23	3	6	225
CMC Yukon 1500 SLE/SUV	16	19	4.8	8	285
Cadillac SRX V8/SUV	16	21	4.6	8	320
Nissan Pathfinder SE/SUV	16	21	3.5	6	240
Chevrolet TrailBlazer LT/SUV	16	21	4.2	6	275
BMW X5 4.4i/SUV	16	22	4.4	8	325
Honda Pilot LX/SUV	17	22	3.5	6	240
Acura MDX/SUV	17	23	3.5	6	265
Mitsubishi Endeavor XLS/SUV	17	21	3.8	6	215
Isuzu Rodeo S/SUV	17	21	3.2	6	193
Nissan Xterra XE V6/SUV	17	20	3.3	6	180
Lexus RX 330/SUV	18	24	3.3	6	230
Toyota 4Runner SR5 V6/SUV	18	21	4	6	245
Ford Escape XLS/SUV	18	23	3	6	201
Toyota Highlander V6/SUV	18	24	3.3	6	230
Suzuki XL-7 EX/SUV	18	22	2.7	6	185
Land Rover Freelander SE/SUV	18	21	2.5	6	174
Pontiac Aztekt/SUV	19	26	3.4	6	185
Buick Rendezvous CX/SUV	19	26	3.4	6	185
Chevrolet Tracker/SUV	19	22	2.5	6	165
Hyundai Santa Fe GLS/SUV	20	26	2.7	6	173
Jeep Liberty Sport/SUV	20	24	2.4	4	150
Honda Element LX/SUV	21	24	2.4	4	160
Honda CR-V LX/SUV	21	25	2.4	4	160
Saturn VUE/SUV	21	26	2.2	4	143
Mitsubishi Outlander LS/SUV	21	27	2.4	4	160
Mazda Tribute DX 2.0/SUV	22	25	2	4	130
Toyota RAV4/SUV	22	27	2.4	4	161

Figure 7: Analysis for SUV Category

From figure 7, which indicates the analysis of the SUV Category, it was discovered that for:

Retail Price and Dealer Cost: Mercedes-Benz G500/SUV ranked the topmost with the highest retail price of £76, 870 and Dealer Cost of £71, 540, while Honda Element LX/SUV ranked the lowest on retail price of £18, 690 and Dealer Cost of £17, 334 respectively.

Engine Size, Miles per gallon and Cylinder size: It was discovered that Ford Excursion 6.8 XLT/SUV that have 10 cylinders ranked the topmost with an engine size of 6.8 liters however, the Hwy MPG and City MPG were not reported. A 4 cylinder Toyota RAV4/SUV with Engine Size of 2.4 liters ranked lowest in terms of miles per gallon in that it has the highest highway miles per gallon of 27 and city miles per gallon of 22 respectively.

2.2.4 Correlation Analysis (Non Visual)

Correlations are useful for describing simple relationships between different data attributes. It's a measure of to what extent does a change in one variable affect the other variable. Correlation analysis of the car dataset was carried out using the python programming language.

```
import numpy as np
```

```
df.corr()
```

	Sedan	Sports Car	SUV	Wagon	Minivan	Pickup	AWD	RWD	Retail Price	Dealer Cost	Engine Size (l)	Cyl	HP
Sedan	1.000000	-0.416041	-0.467207	-0.317670	-0.256178	-0.282015	-0.318038	-0.096914	-0.176489	-0.168657	-0.301108	-0.174956	-0.254785
Sports Car	-0.416041	1.000000	-0.145188	-0.098718	-0.079609	-0.087638	-0.098833	0.393055	0.381856	0.376644	0.079924	0.058811	0.342155
SUV	-0.467207	-0.145188	1.000000	-0.110859	-0.089400	-0.098416	0.411247	-0.237484	0.041928	0.036907	0.263747	0.197042	0.112163
Wagon	-0.317670	-0.098718	-0.110859	1.000000	-0.060786	-0.066917	0.056840	-0.014875	-0.055653	-0.052491	-0.105805	-0.080575	-0.083742
Minivan	-0.256178	-0.079609	-0.089400	-0.060786	1.000000	-0.053963	-0.035008	-0.104884	-0.056789	-0.058540	0.067637	0.016979	-0.035286
Pickup	-0.282015	-0.087638	-0.098416	-0.066917	-0.053963	1.000000	0.169126	0.135531	-0.098371	-0.102325	0.194238	0.071321	0.030395
AWD	-0.318038	-0.098833	0.411247	0.056840	-0.035008	0.169126	1.000000	-0.307756	0.099985	0.094665	0.203412	0.142591	0.140110
RWD	-0.096914	0.393055	-0.237484	-0.014875	-0.104884	0.135531	-0.307756	1.000000	0.403593	0.403410	0.264899	0.299100	0.382689
Retail Price	-0.176489	0.381856	0.041928	-0.055653	-0.056789	-0.098371	0.099985	0.403593	1.000000	0.999132	0.571753	0.628757	0.826945
Dealer Cost	-0.168657	0.376644	0.036907	-0.052491	-0.058540	-0.102325	0.094665	0.403410	0.999132	1.000000	0.564498	0.624204	0.823746
Engine Size (l)	-0.301108	0.079924	0.263747	-0.105805	0.067637	0.194238	0.203412	0.264899	0.571753	0.564498	1.000000	0.897564	0.787435
Cyl	-0.174956	0.058811	0.197042	-0.080575	0.016979	0.071321	0.142591	0.299100	0.628757	0.624204	0.897564	1.000000	0.775799
HP	-0.254785	0.342155	0.112163	-0.083742	-0.035286	0.030395	0.140110	0.382689	0.826945	0.823746	0.787435	0.775799	1.000000

Figure 8: Correlation Analysis Table for the Car Attributes

Correlations are defined with a unit-free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by r. Statistical significance is indicated with a p-value. Therefore, correlations are typically written with two key numbers: r = and p = . The closer r is to zero, the weaker the linear relationship.

Positive r values indicate a positive correlation, where the values of both variables tend to increase together. Negative r values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease. The p-value gives an evidence that can give a meaningful conclusion that the population correlation coefficient is likely different from zero, based on what we observe from the sample. Leishi 2023.

Below are the definition of different correlation (i.e. r)values:

a. When r = 1, there exists a Perfect Positive Correlation
b. When r = 0.9, there exists a High Positive Correlation
c. When r = 0.5, there exists a Low Positive Correlation
d. When r = 0, there exists a No Correlation
e. When r = -0.5, there exists a Low Negative Correlation
f. When r = -0.9, there exists a High Negative Correlation
g. When r = -1, there exists a Perfect Negative Correlation

Based on the general correlation carried out on the whole data, it was discovered that others exists stronger correlation between some variables in comparison to other, hence below is the list of selected variable with stronger correlation.

a. Retail Price
b. Dealer Cost
c. HP (Horse Power)
d. Hwy MPG (Highway Miles per gallon)
e. City MPG (City Miles per gallon)
f. Cylinder
g. Engine Size

2.2.5 Correlation between Retail Price and Other Variables ((Dealer Cost, HP, Cyl and Engine Size)

```
# Correlation between Retail Price and Other Variables (Dealer Cost, HP, Cyl and Engine Size)

[13]: corr = df.corr()
      corr["Retail Price"]["Dealer Cost"]

[13]: 0.9991316206272151

[14]: corr = df.corr()
      corr["Retail Price"]["HP"]

[14]: 0.8269450085627253

[15]: corr = df.corr()
      corr["Retail Price"]["Cyl"]

[15]: 0.6287572874923403

[16]: corr = df.corr()
      corr["Retail Price"]["Engine Size (l)"]

[16]: 0.5717529618466338
```

Figure 9: Correlation Analysis of Retail Price and Other Variables

Figure 9 reveals that there exists a High Positive Correlation Coefficients (r) of 0.9 between Retail Price and Dealer Cost and Low Positive Correlation coefficients (r) of 0.8, 0.6 and 0.5 between dealer cost and HP, Cyl and engine size (l) respectively.

2.2.6 Correlation between Dealer Cost and Some Other Variables (HP, Cyl, and Engine Size)

```
# Correlation between Dealer Cost and Some Variables (HP, Cyl, and Engine Size)

[19]: corr = df.corr()
      corr["Dealer Cost"]["HP"]

[19]: 0.8237464987316104

[22]: corr = df.corr()
      corr["Dealer Cost"]["Cyl"]

[22]: 0.6242044223270925

[23]: corr = df.corr()
      corr["Dealer Cost"]["Engine Size (l)"]

[23]: 0.5644979779575726
```

Figure 10: Correlation between Dealer Cost and Some Other Variables

From figure 10, it showed that there exist a High Positive Correlation coefficients (r) of 0.8 between Dealer Cost and HP and Low Positive Correlation coefficients (r) of 0.6 between Dealer Cost and Cylinder respectively while a Low Positive Correlation coefficient (r) of 0.5 exist between Dealer Cost and Engine size (l).

2.2.7 Correlation between Cylinder (Cyl) and Some Variables (Engine Size and HP)

```
# Correlation between Cyl and Some Variables (Engine Size and HP)

[24]: corr = df.corr()
      corr["Cyl"]["Engine Size (l)"]

[24]: 0.8975637090473843

[26]: corr = df.corr()
      corr["Cyl"]["HP"]

[26]: 0.7757985325285554
```

Figure 11: Correlation between Cylinder and Some Other Variables

From figure 11, there exists High Positive Correlation coefficients (r) of 0.9 and 0.8 between Cyl and HP and Cyl and Engine Size (l) respectively.

2.2.8 Correlation between HP and Engine Size

```
# Correlation between HP and Engine Size

[77]: corr = df.corr()
      corr["HP"]["Engine Size (l)"]

[77]: 0.7874349451231397
```

Figure 12: Correlation Analysis between HP and Engine Size

From figure 12, a High Positive Correlation coefficient (r) of 0.7 exists between HP and Engine size.

3.0 Good and Bad Visualizations (Activity 3.1)

The tasks under this activity involves the:

a. Identification of three Good and Bad visuals each.
b. Writing down why they are bad or good.
c. Suggesting better alternatives and possible improvements.

3.1.1 Identification of 3 Bad Visuals

Below are examples of 3 Bad visuals identified with the sources from where they are collected.

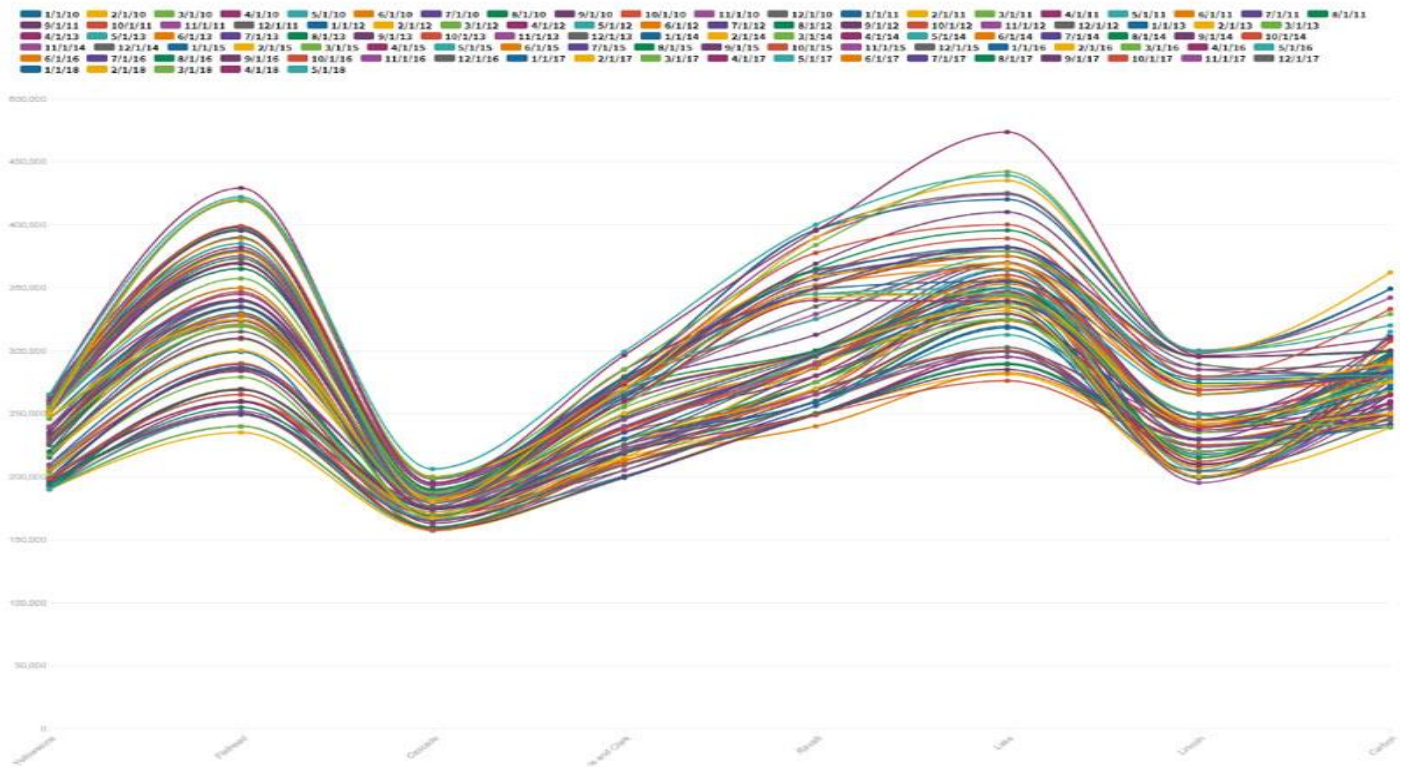


Figure 13: Bad Visual 1- Misleading Data Visualization Examples to Stay Away From by Milan J.

Figure 13 is a graph which represents the number and range of students exam over a period of time. It was a data from students admission into university from minority group and are of lower income. Inspection showed that within these groups the average scores increased.

This visual is categorised as bad visual because:

a. It contains too many lines which makes it difficult to see clearly the information on the visual;
b. The legend does not reflect any information which is as a result of the too many

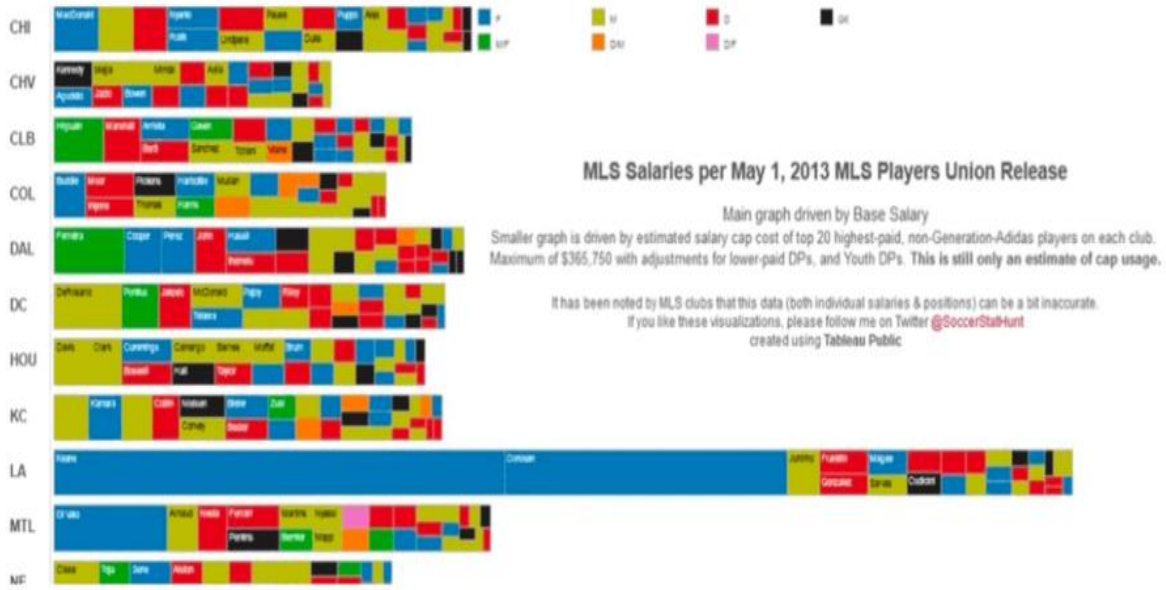


Figure 14: Bad Visual 2- Bad Data Visualization Examples: Mohiuddin O., 2022

Figure 14 is about the MLS salaries per May 1, 2013 by MLS player release. It's considered a bad visual because of the following reasons:

a. It looks too busy with a lot of information
b. It will overload the audience with too many information.
c. It will not permit the audience to get anything of substance.

The above graph can be made better by categorising the MLS salaries into different groups (range) and then use a Bar chart to visualize each group.

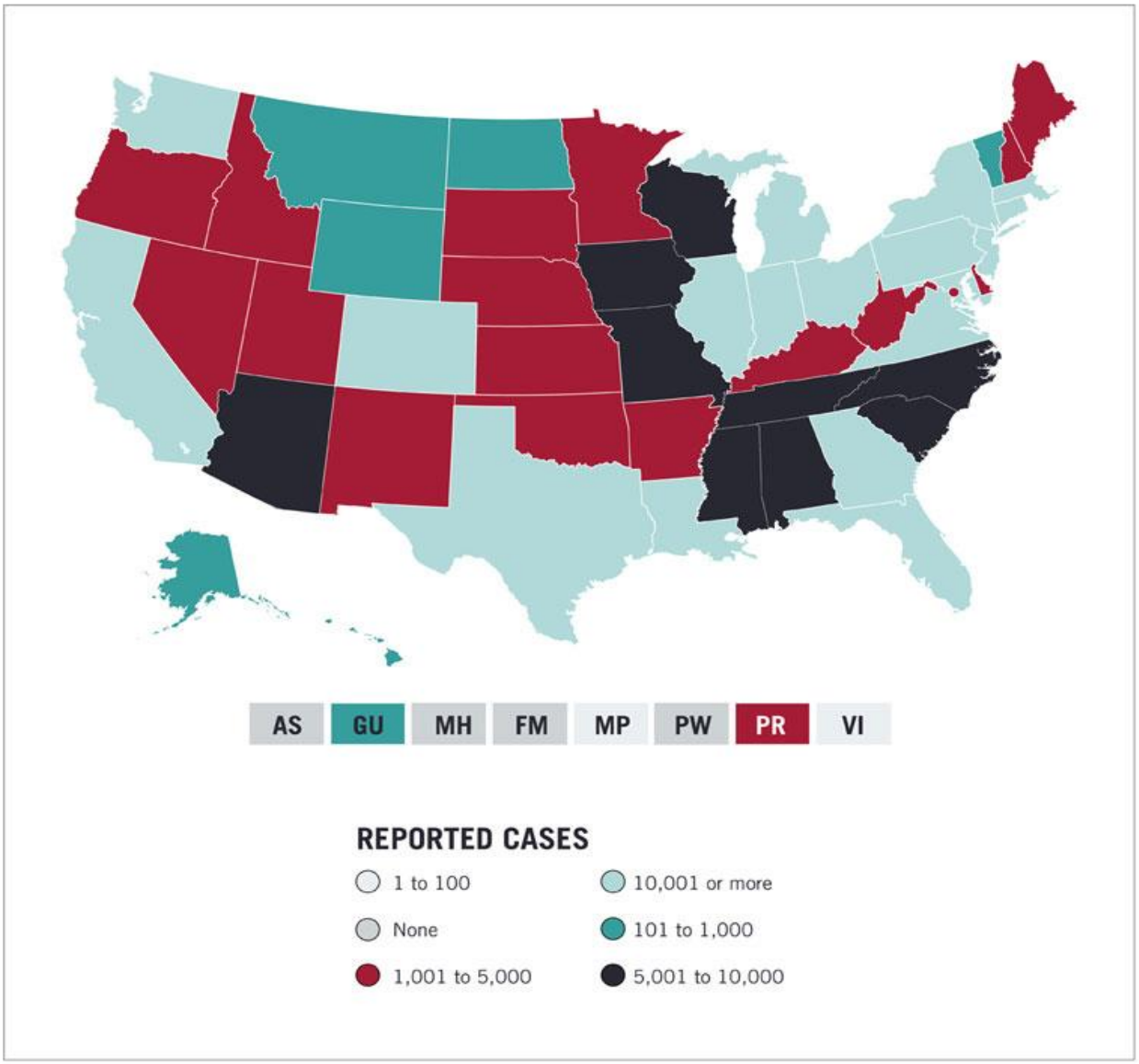


Figure 15: Bad Visual 3: Misleading Data Visualization Examples to Stay Away From by Milan J.

Figure 15, is a graph showing the map of the United States with infection rate. While it is good to use map as visual, this is categorized as a bad visual because:

a. It shows the rate with different colours but it does not use concentration to differentiate between the countries that have the same colour shade.
b. The result of the visual is very confusing and may be misleading to the viewers.

The above visual can be made better by using the same visual type but include different color concentration to show the rate of infection i.e. the deeper the concentration, the higher the rate of infection and vice versa.

3.1.2 Identification of 3 Good Visuals

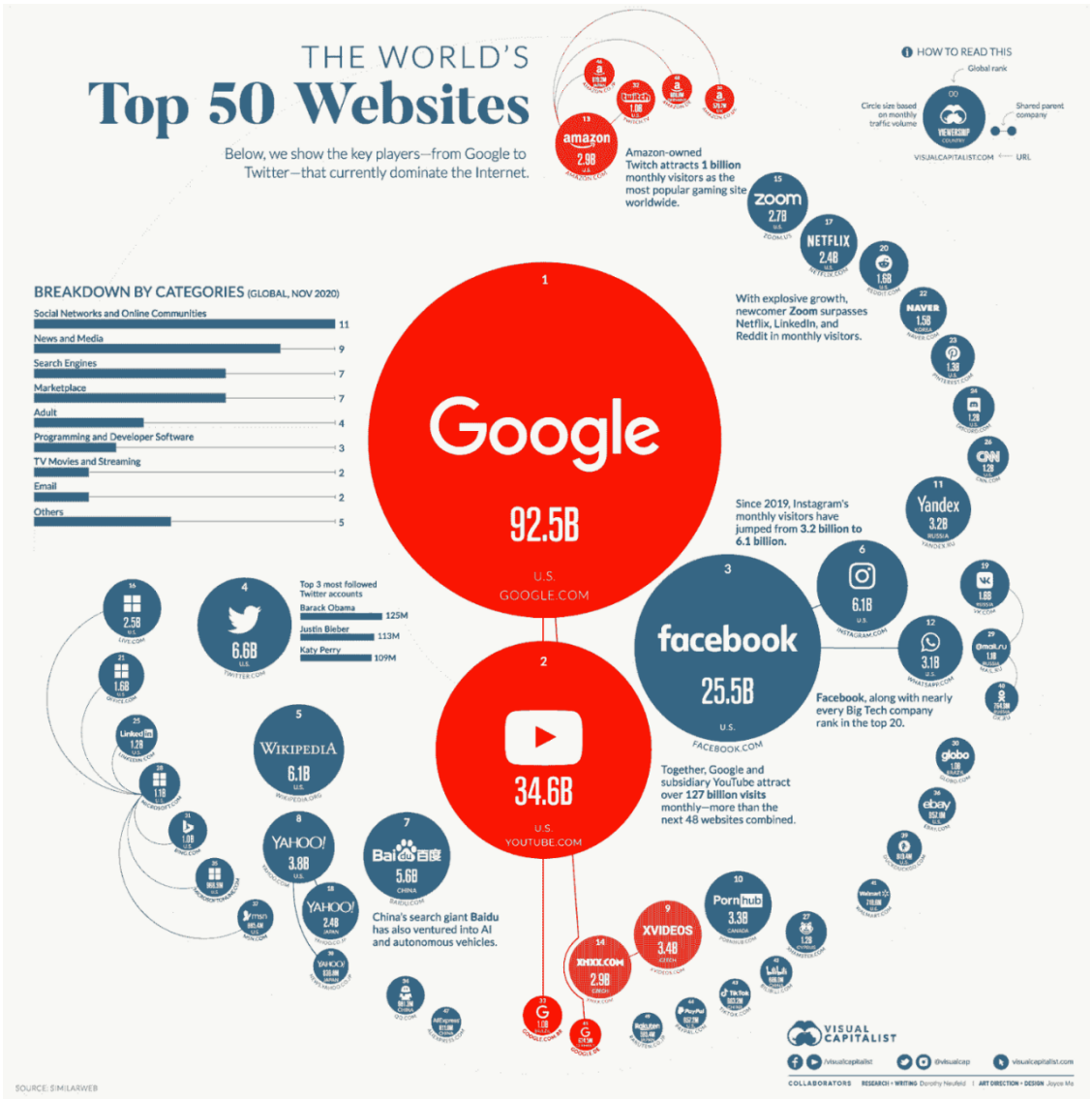


Figure 16: Good Visual 1: The World's Top 50 Websites: Mohiuddin O., 2022.

Figure 16 is a visual representation of the World's top 50 websites. It's categorized as a good visual because:

a. It shows the website with the most traffic with the use of sizes of the bubble charts
b. The sizes of the bubble conveys the relative sizes of different site audiences
c. It includes a smaller bar chart in a larger one to show the industries with the top niche.

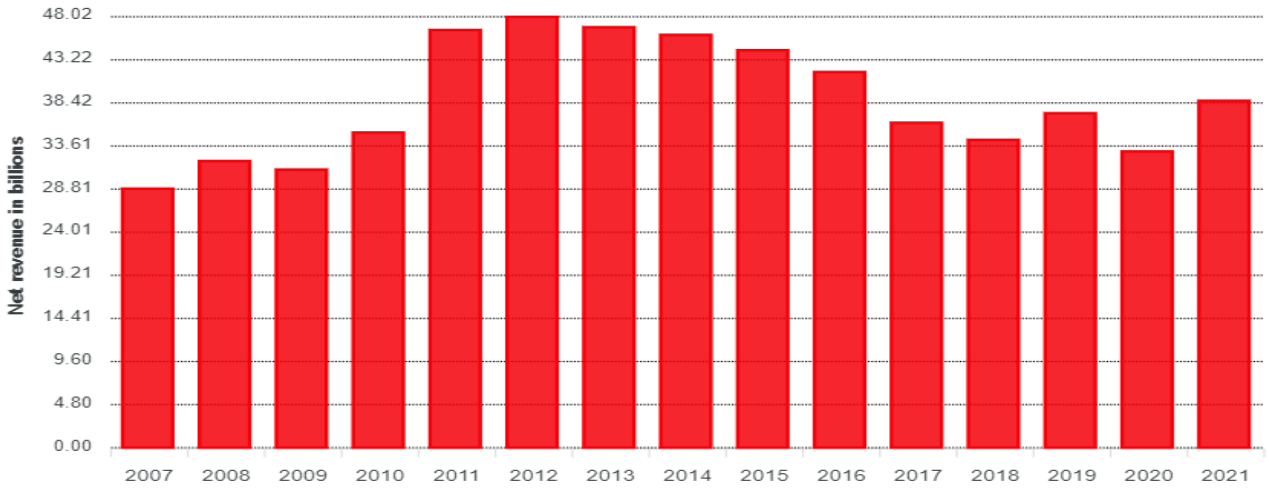


Figure 17: Good Visual 2: Comparison of Coca-Cola Net Revenue Over the years

Figure 17 is a bar chart explaining Coca-Cola's net revenue over the years. This is categorized as a good chart for the following reasons:

a. The vertical axis (y axis) runs from 0 to 48 billion. This satisfies the rule for a bar chart i.e. the y axis was not truncated, it starts from zero.
b. It compares the net revenue that Coca-Cola had over the past years. This allows for the real comparison of the difference between the net revenue between each year.

Figure 18: Good Visual 3: Diets around the World - Mohiuddin O., 2022

Figure 16 is the visual representing diets around the world by location and date. It's interactivity may not be accessible but it's categorized as a good visual for the following reasons:

a. It's a doughnut chart (middle) which visualizes different what people are consuming all over the world in different period.
b. The visual page contains charts for individual countries for what they consume per time.
c. It's well-descriptive in that it shows to a viewer the description of each countries daily diet per time, even without any description.

3.2 Activity 3.2 - Identification of Marks and Channels

This activity involves the re-evaluation of the previously identified Bad and Good visuals in the section 3.1 of this portfolio. It requires the identification of:

a. Analysis of the marks and channels used in each visualization.
b. Critical review of the effectiveness and expressiveness of the visualizations based on the principles introduced in the class module.

3.2.1 Marks and Channels for Bad Visuals

a. Bad Visual 1- Misleading Data Visualization Examples to Stay Away From by Milan J.

The mark used in this visual was Line (spatial dimension is 1D), and the channel was coloured vertical lengths and horizontal positions. The mark line visual was a good one but it has too many lines thereby leading to confusion on the part of the viewer in identifying each attribute. In addition, it does not allow the viewer to differentiate bewteen each of the lines thereby loosing details on the important data points. No clear data points can be seen relating to the number and range of students' exams over a period of the time in question. A bar chart would be a better option in the regard because it will specify each attirbute and separability and pop out of each of the attributes from each other.

b. Bad Visual 2- Bad Data Visualization Examples: Mohiuddin O., 2022

The mark used in this visual was Area (which has a spatial dimension of 2D), and the channel was colour in vertical length and horizontal position. The visual looks too busy with lots of information which will overload the audience with too many information. It does not permit the audience to get information of substance. Conclusively, the channel lacks accuracy, discriminability, separability and pop out. The visual can be made better by categorising the MLS salaries into different groups (range) and then use a Bar chart to visualize each group. The colour channel can be retained to add to the separability, discriminability and pop out of each category and information being sent to the audience.

c. Bad Visual 3: Misleading Data Visualization Examples to Stay Away From by Milan J.

The mark used in this visual was Area (which has a spatial dimension of 2D) and the channel was colour. It shows the infection rate with different colours in different countries but it lost the ability to differentiate the countries with highest or lowest infection rate. The result of the visual is very confusing and may be misleading to the viewers. Consequently, the visual lacks expressiveness (channel ranking) and effectiveness (Accuracy, Separability and Pop out). The visual can be made better by using the same visual type but include different color concentration to show the rate of infection (channel ranking) i.e. the deeper the

concentration, the higher the rate of infection and vice versa. This will bring out the separability in terms of the sizes and the pot out will be noticed by the colour intensity.

3.2.2 Marks and Channels for Good Visuals

a. Good Visual 1: The World's Top 50 Websites: Mohiuddin O., 2022.

The visual is a combination of bar chart (with points as mark and vertical lengths and horizontal position as channels) and bubble charts with point as the Marks and channel used were colour and size. The visual satisfied the effectiveness (Accuracy, Separability and popout conditions of an effective visual) i.e. the important items which are World's most visited website (Google, YouTube, Facebook, etc) were made most noticeable by their sizes as the first 3 most noticeable websites. Considering the expressiveness (order) were seen as ordered on the visual.

b. Good Visual 2: Comparison of Coca-Cola Net Revenue Over the years.

The mark used in this visual was a column chart with line mark and the channel was colour with vertical lengths and horizontal position. This visual satisfies all the requirements of effectiveness and expressiveness of a visual. It has separability and accuracy. The y axis was not truncated in that it starts from up to 48 billion.

c. Good Visual 3: Diets around the World - Mohiuddin O., 2022

The mark used in this visual was point and the channel was colour. It has separability (by separating each countries on different charts to reflect the daily diet consumption per country) and accuracy in that each doughnut chart adds up to 100%.

4. 1 Visual Analysis of Nominal, Ordinal and Numeric Data Using Knime (Activity 4.1)

Different data types requires different type of analyses hence different techniques to achieve best result. This activity involves the use of Knime software to carry out visual data analytics on a dataset of choice (car dataset) which must contain nominal, ordinal and numeric data.

4.1 Nominal Data:

It involves the following tasks:

41.1. Visual analytics tasks that can be carried out on the car dataset

Based on the car dataset, the following visual analytics tasks can be can be carried out:

a. Importation of dataset into knime: The car dataset was imported into knime using the csv reader node.
b. Data Cleansing: The tasks that were carried out in cleansing the data were conversion of some attributes to integers. Some data were in the wrong data type, the string to number node was used to overcome this problem. The data contains some missing values, these was sorted by using the missing value node to correct this anomaly. For strings data type, missing values were configured to be replaced with most frequent values while for number (integers and double) were replaced with the mean values.
c. Data Reduction: Selection of target attributes was done by using the column filter node. The non-relevant attributes were removed leaving the relevant column.
d. Descriptive Analysis: Some descriptive statistics were done on the dataset by using the statistics node. The data explorer node was applied to further extend the visualization of the descriptive statistics and create an interactive view.
e. Visualization: Visualization of the outcome some of some descriptive statistics (numerical data), nominal and categorical data were done by using the Bar Chart node, Bar Chart (Labs) Node, Scatter Plot (Labs) and Pie Chart (Labs) node to visualize the data.
f. Correlation Analysis: Correlation analysis was used to detect relationship between the attributes. Two nodes i.e. Linear and Rank correlation nodes were used to do the correlation analysis.
g. Visualization Analysis of Correlation: The outcome of the correlation analysis was visualised using basically scatter plot node.
h. Data Partitioning: Data partitioning was done by using the partitioning node. The filtered dataset was partitioned using relative method into 70% for training and 30% for test datasets. The method used for the sampling was random selection.
i. Regression Analysis: Regression analysis is useful for prediction. In this case, the attribute predicted was Retail Price using the Random Forest Learner and Random Forest predictor respectively.
j. Model Performance: The model used for prediction was subjected to scoring by using the node scorer. This helps to know the performance of our model to learn from the 70% training data and to predict the retail price using the 30% test dataset.

4.1.2 Extension or redesigning the Knime workflow to support the specified visual analysis tasks.

Below is a picture of the redesigning of the workflow in knime to support the specified visual analysis tasks.

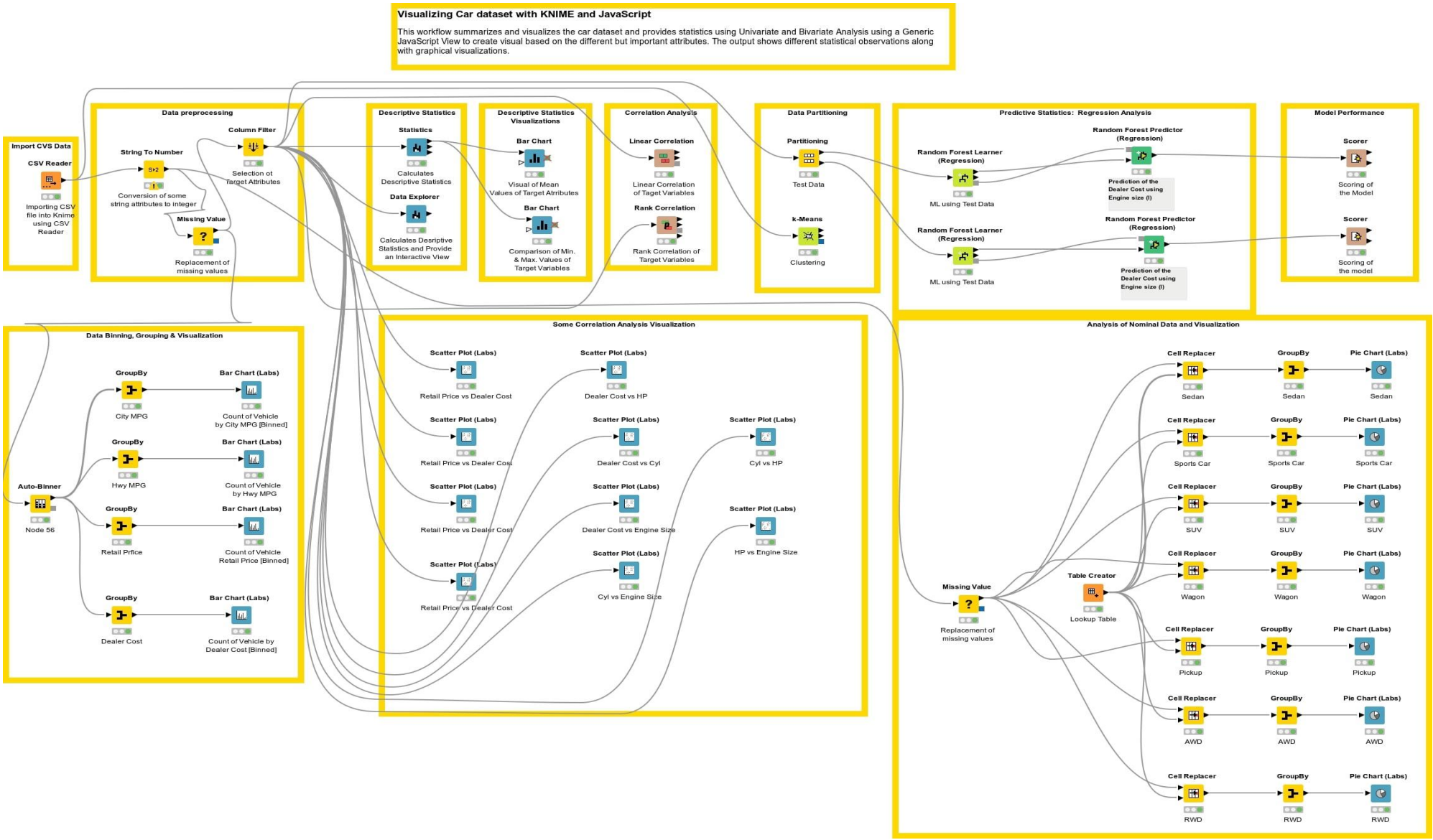


Figure 19: A Representation of the Knime Workflow Design for Car Dataset.

4.2 Result Findings from the Analysis of the Car Dataset Using Knime (Activity 4.2)

Knime workflow was designed to analyse, visualize and summarize the car dataset and provide statistics using Univariate and Bivariate Analysis using knime node and a generic JavaScript View to create visual based on the different but important attributes. The output shows different statistical observations along with graphical visualizations. Below are summary with screen shot of the workflow(s) and visualizations, and explanation of the findings.

4.2.1 Descriptive Statistics (Numeric Attributes)

The data explorer node was applied to further extend the visualization of the descriptive statistics and create an interactive view. Below are the results of the car dataset descriptive statistics.

a. Below is the table showing the general statistics of the numeric target datasets.

Numeric | Nominal | Data Preview

Show 10 entries Search:

Column	Exclude Column	Minimum	Maximum	Mean	Median	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros
Retail Price	☐	10280	192465	32774.855	27635	19431.717	377591612.888	2.798	13.879	14027638	0
Dealer Cost	☐	9875	173560	30014.701	25294.500	17642.118	311244318.716	2.835	13.946	12846292	0
Engine Size (l)	☐	1.300	8.300	3.197	3	1.109	1.229	0.708	0.542	1368.200	0
Cyl	☐	-1	12	5.776	6	1.623	2.633	0.234	1.397	2472	0
HP	☐	73	500	215.886	210	71.836	5160.415	0.930	1.552	92399	0
City MPG	☐	10	60	20.089	19	5.127	26.285	2.981	17.322	8598.251	0
Hwy MPG	☐	12	66	26.906	26	5.603	31.390	1.374	6.765	11515.681	0

Figure 20: General Descriptive Statistics of Target Attributes

Figure 20 shows the Minimum, maximum, Mean, Median, Standard Deviation, Variance and sum overall for all the target numeric datasets. Ultimately, the main aim of this report is to predict the Retail Price of the vehicles. The retail price has a maximum value of 192,465, minimum value of 10,280 and mean value of 32, 774.9 respectively.

Some visualizations were done using the data explorer node. Below are some of the visuals from the descriptive statistics.

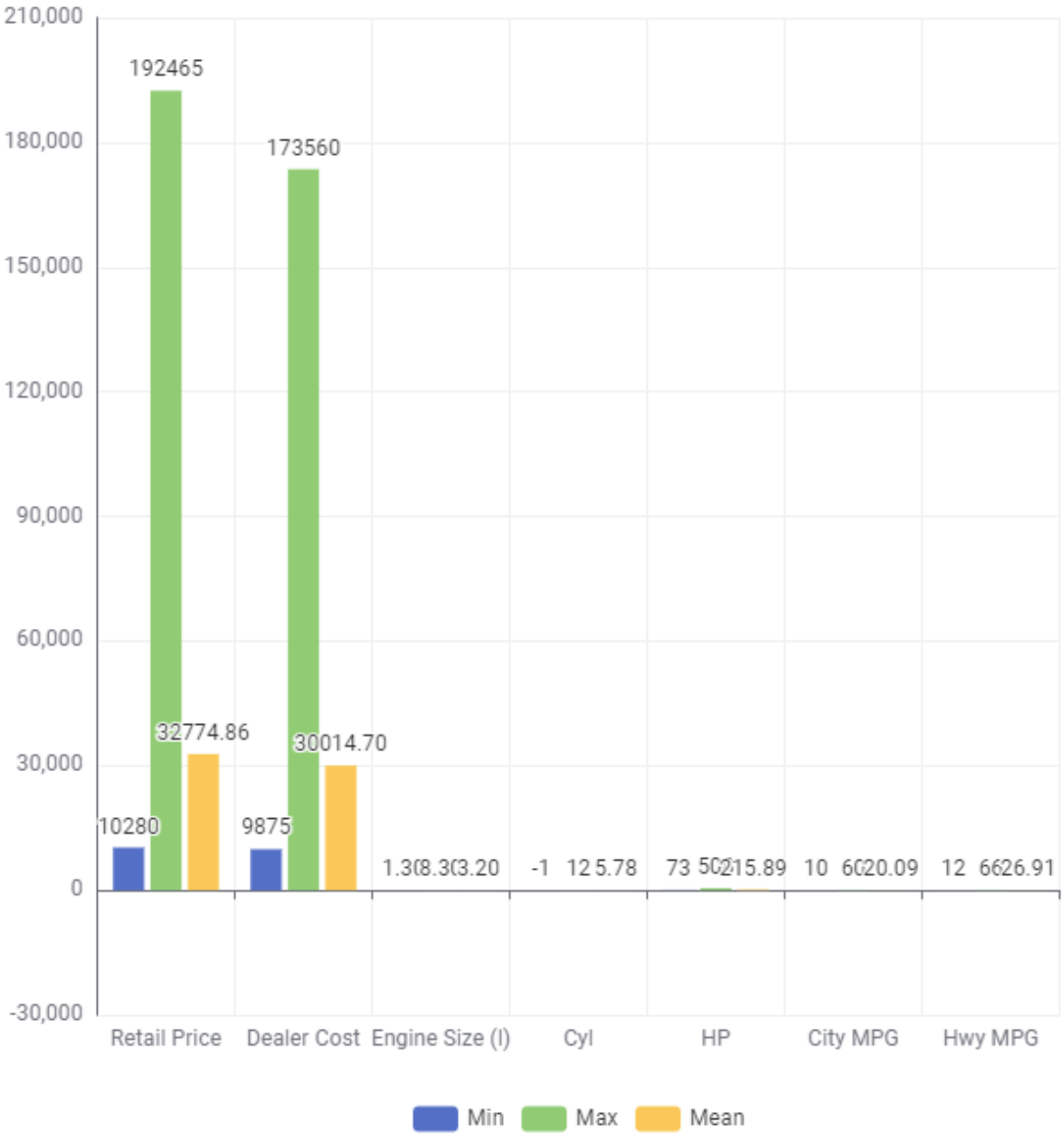


Figure 21: Comparison of Minimum, Maximum and Mean Values of Target Attributes

From figure 21, it can be seen that there is little difference between the maximum values of retail price and dealer costs, hence the mean values had little difference as well.

4.3 Correlation Analysis

Linear correlation analysis between a variable x and another variable y is a measure of how similar their deviations from their respective means are. It is a measure of the relationship between two variables i.e. measure of association. The correlation analysis was carried out between the target variables using the linear correlation node.

Correlation measure - 3:37 - Rank Correlation (Rank Correlation of)

File Edit Hilite Navigation View

Table "default" - Rows: 21 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	First column name	Second column name	Correlation value	p value	Degrees of freedom
Row0	Retail Price	Dealer Cost	0.9990560603859743	0.0	426
Row1	Retail Price	Engine Size (l)	0.669441191029097	0.0	426
Row2	Retail Price	Cyl	0.7252735321148881	0.0	426
Row3	Retail Price	HP	0.8659515903362897	0.0	426
Row4	Retail Price	City MPG	-0.6769093841193561	0.0	426
Row5	Retail Price	Hwy MPG	-0.5816959980499623	0.0	426
Row6	Dealer Cost	Engine Size (l)	0.6609961995966402	0.0	426
Row7	Dealer Cost	Cyl	0.7174789311064117	0.0	426
Row8	Dealer Cost	HP	0.8625829244665417	0.0	426
Row9	Dealer Cost	City MPG	-0.6723251720436262	0.0	426
Row10	Dealer Cost	Hwy MPG	-0.5785143606100563	0.0	426
Row11	Engine Size (l)	Cyl	0.9251361834847265	0.0	426
Row12	Engine Size (l)	HP	0.8077416354075443	0.0	426
Row13	Engine Size (l)	City MPG	-0.8196264785719536	0.0	426
Row14	Engine Size (l)	Hwy MPG	-0.7309020218524794	0.0	426
Row15	Cyl	HP	0.8058056983361247	0.0	426
Row16	Cyl	City MPG	-0.7801000184762394	0.0	426
Row17	Cyl	Hwy MPG	-0.6820206939223795	0.0	426
Row18	HP	City MPG	-0.7550863639836727	0.0	426
Row19	HP	Hwy MPG	-0.6660285331076351	0.0	426
Row20	City MPG	Hwy MPG	0.9271933100717522	0.0	426

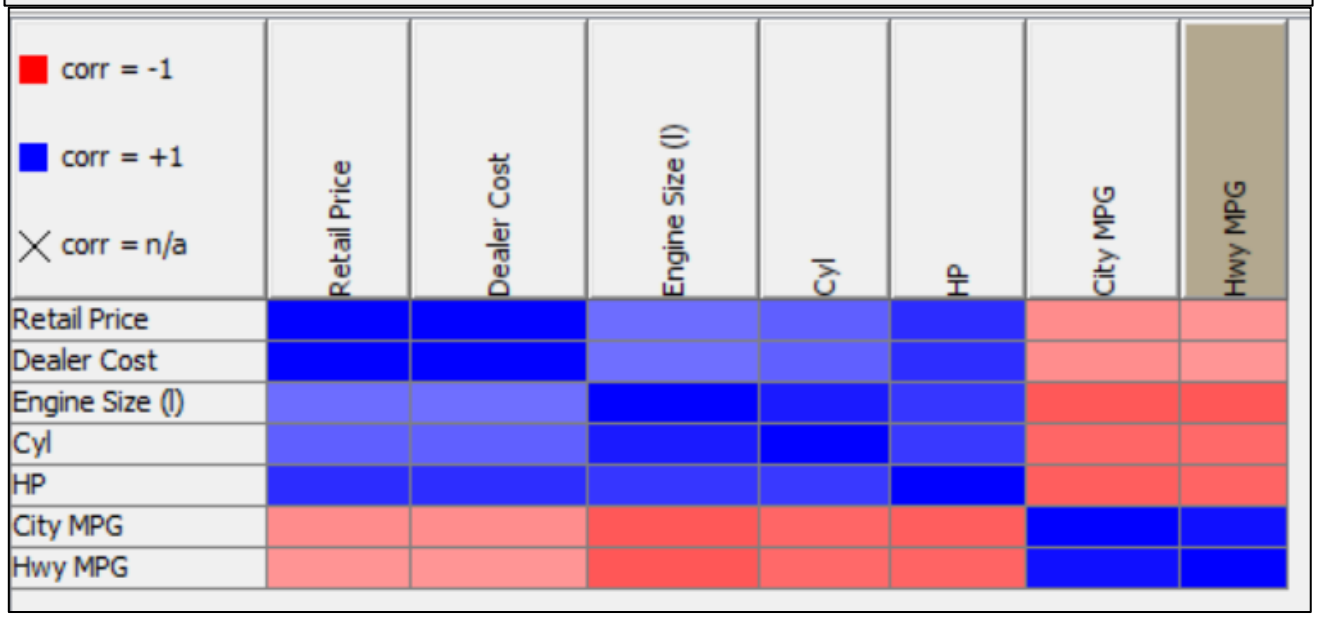


Figure 22: Correlation Analysis between Target Variables

From figure 22, it was discovered that there exists positive and negative correlation between one variable or the other. A further pair to pair correlation analysis based on the general correlation analysis result and based on target variables are described and explained below.

4.3.1 Correlation Analysis Between Retail Price and Some Other Variables (Visual Representation)

a. Retail Price and Dealer Cost

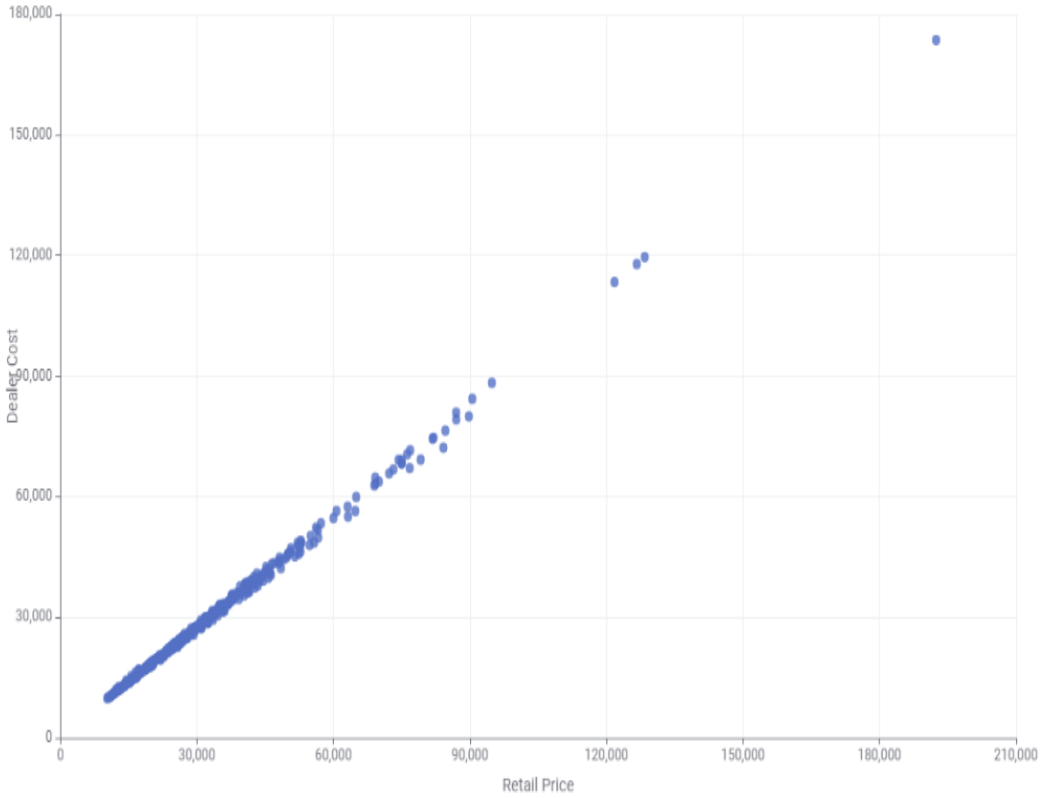


Figure 23: Correlation Between Retail and Dealer Cost

The non visual analysis of the correlation between retail price and dealer cost gave a correlation coefficient (r) of 1. This was further supported visualy by figure 23 which reveals that there exists a High Positive linear correlation between Retail Price and Dealer Cost i.e. as variable Retail price increase, variable dealer cost also increase and vice versa.

b. Retail Price and Engine Size (l)

Figure 24: Correlation between Retail Price and Engine Size (l)

From the non-visual analysis, the correlation coefficient (r) between Retail price and Engine Size (l) is 0.5. Figure 24 supports this outcome visually that there exists a Low Positive Correlation between the variables. The relationship lowly positively correlated.

c. Correlation Analysis Between Retail Price and HP

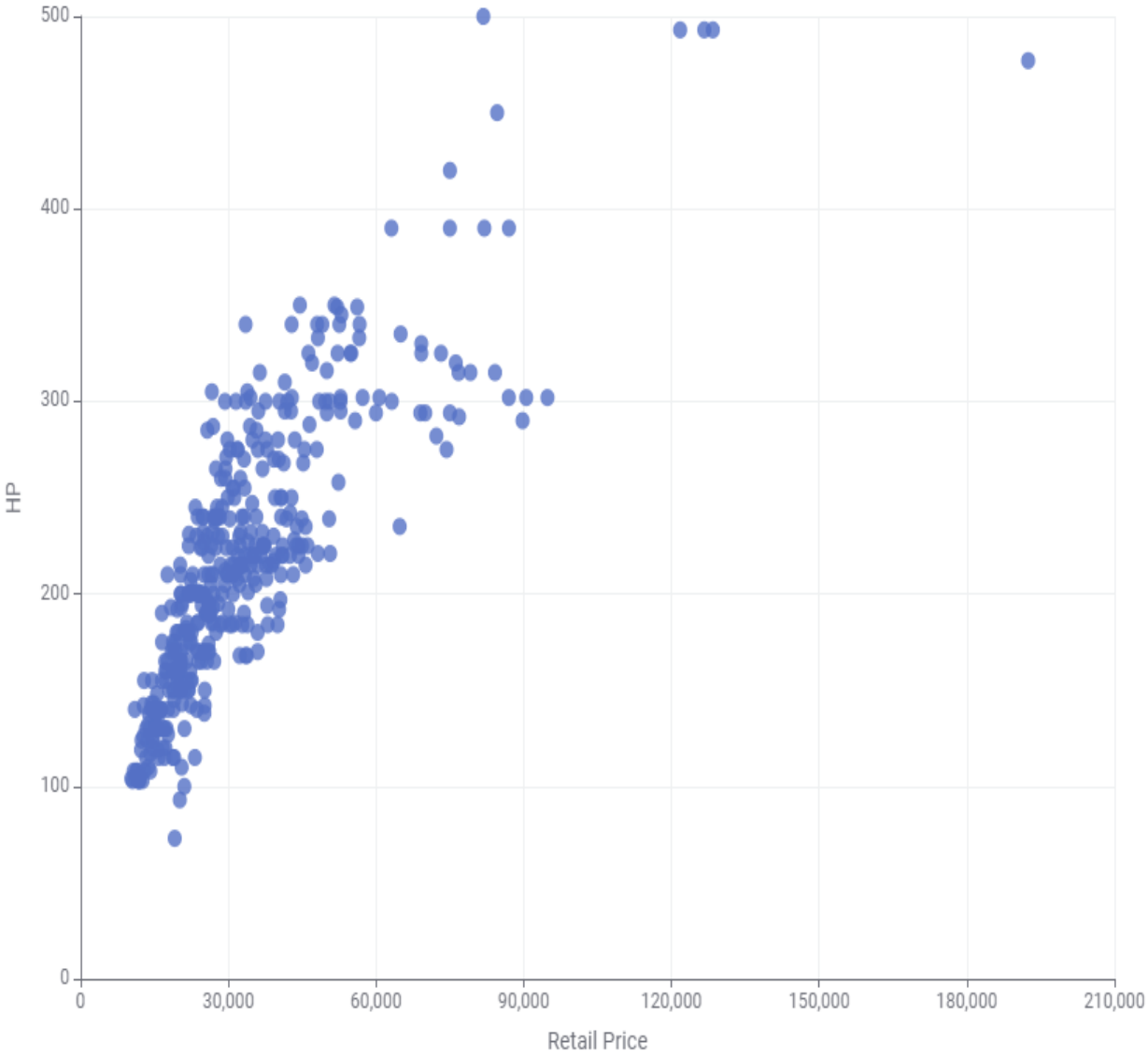


Figure 25: Correlation between Retail Price and HP

The non-visual analysis in the previous section of this report showed that the correlation coefficient (r) between Retail price and HP is 0.8 i.e. a high positive correlation. This is evident in figure 24 which visually supports that there exists a high Positive Correlation between the variables. The relationship is highly positively correlated.

d. Retail price and Cyl (Cylinder)

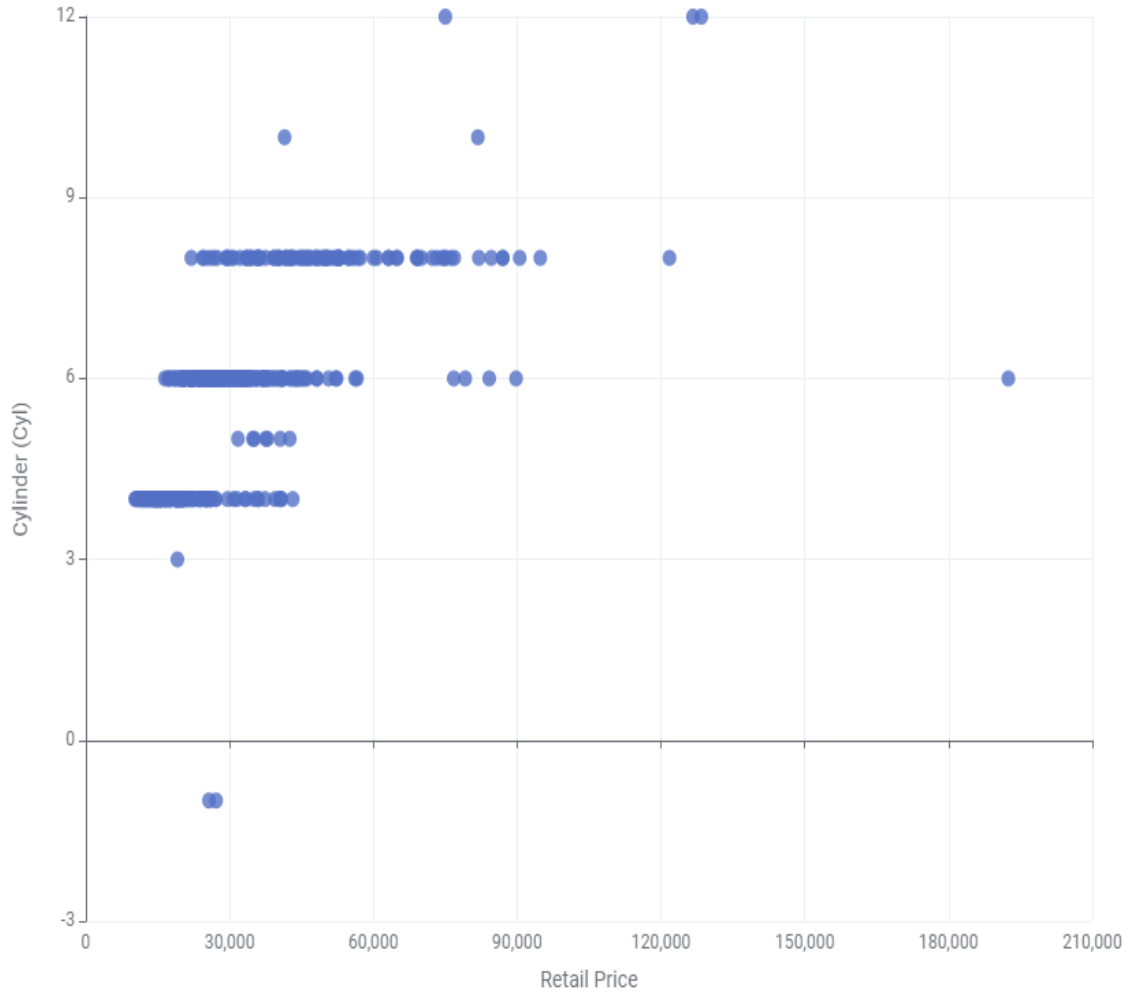


Figure 26: Correlation between Retail Price and Cylinder (Cyl)

From the non-visual analysis, the correlation coefficient (r) between Retail price and Cyl is 0.6. Figure 24 supports this outcome visually that there exists a Low Positive Correlation between the variables. The relationship highly positively correlated.

4.3.2 Correlation Analysis Between Dealer Cost and Some Other Variable (Visual Representation)

a. Dealer Cost vs HP

Figure 27: Correlation Analysis between Dealer Cost and HP

From the non-visual analysis of the dataset, the correlation coefficient (r) between dealer cost and HP was 0.8, this signifies a High Positive Correlation. Figure 27 corroborated the result by the non-visual outcome of the highly positively correlated relationship between the two variables.

b. Dealer Cost vs Engine Size (l)

Figure 28: Correlation Analysis between Dealer Cost and Engine Size (l)

From the non-visual analysis, the correlation coefficient (r) between Dealer and engine size (l) reveals that the correlation coefficient (r) was 0.6 i.e. highly positively correlated. Figure 28 supports this outcome visually. The relationship highly positively correlated.

c. Dealer Cost vs Cyl

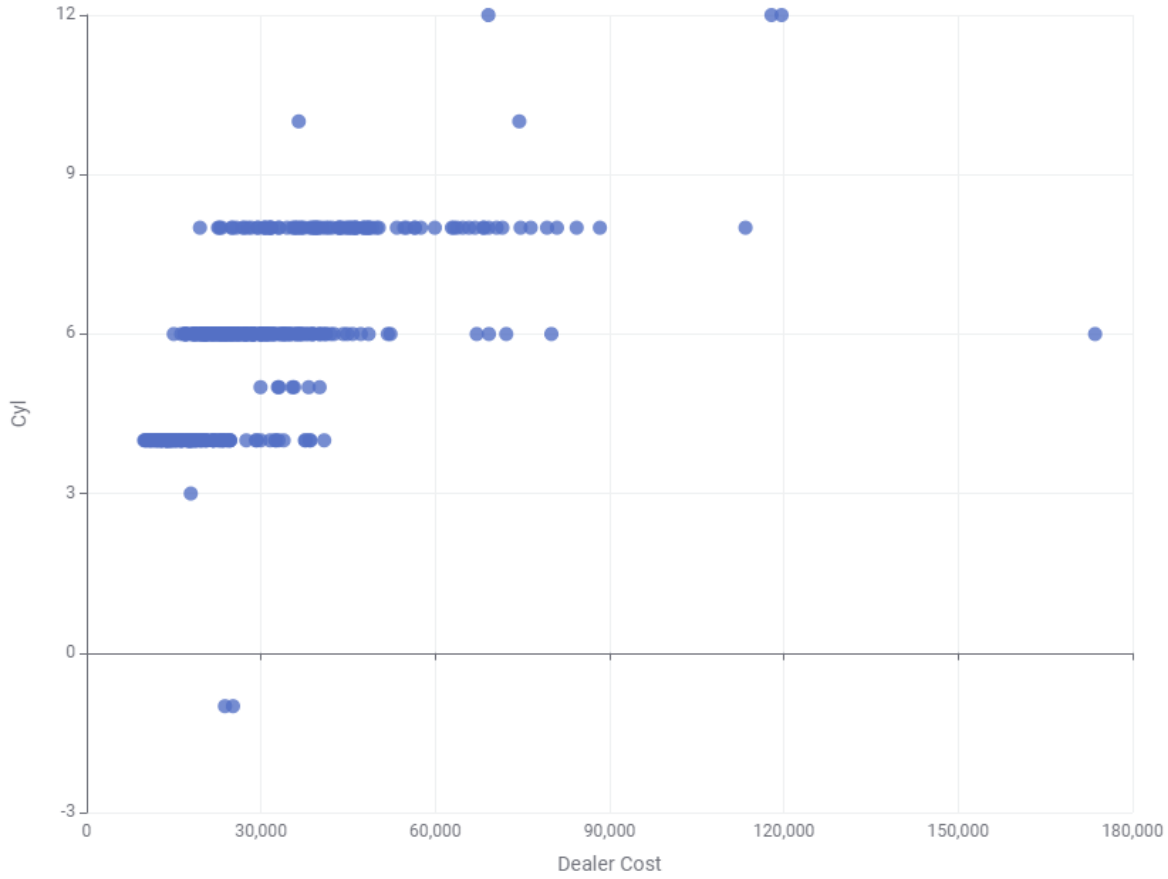


Figure 29: Correlation Analysis between Dealer Cost and Cyl

From the non-visual analysis, the correlation coefficient (r) between Dealer and Cyl reveals was 0.5, this reveal that the relationship between the two variables was lowly positively correlated. Figure 28 supports this outcome visually.

4.3.3 Correlation Analysis Between Engine Size (l) and Some Other Variable (Visual Representation)

a. Engine Size (l) vs HP

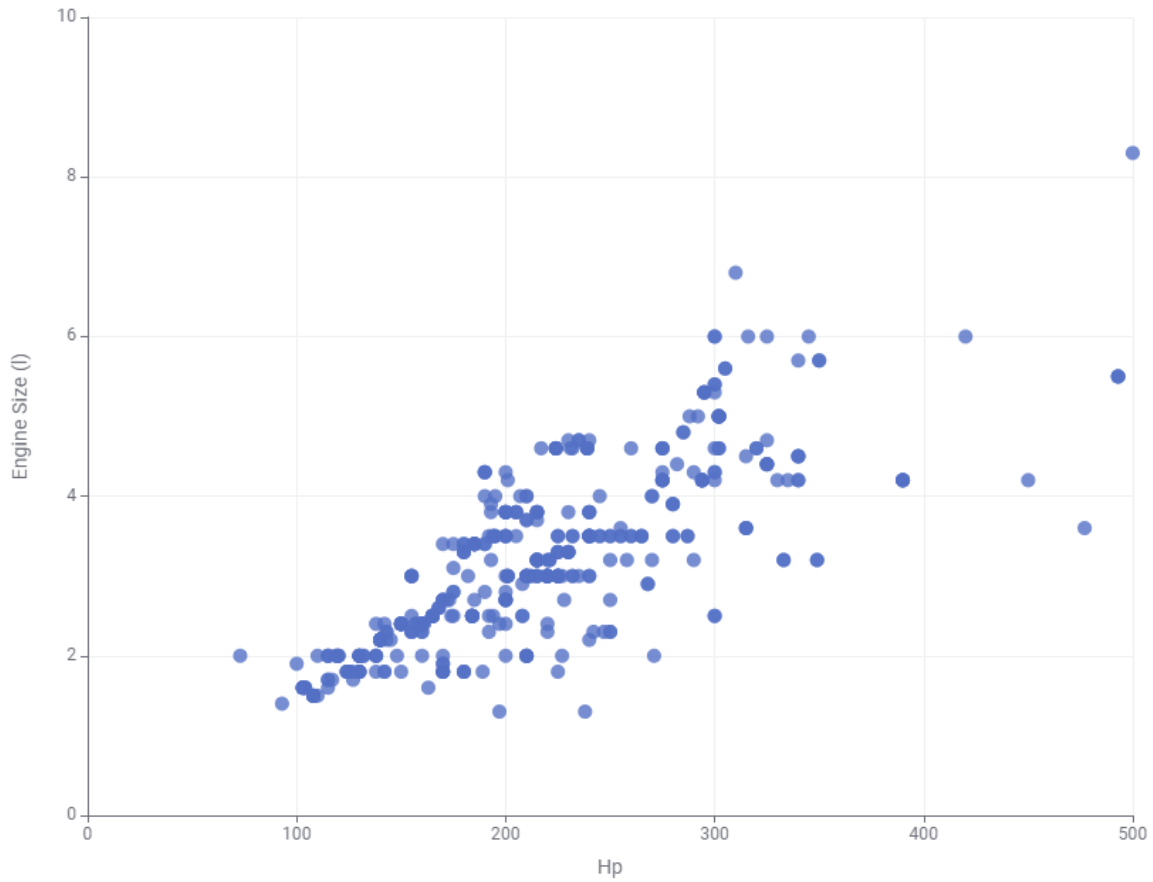


Figure 30: Correlation Between Engine Size (l) and HP

The result of the non-visual correlation analysis between Engine Size (l) and HP reveals coefficient correlation (r) of 0.7 i.e. their relationship is Highly Positively Correlated. This is visually represented in figure 30.

b. Engine Size vs Cyl (Cylinder)

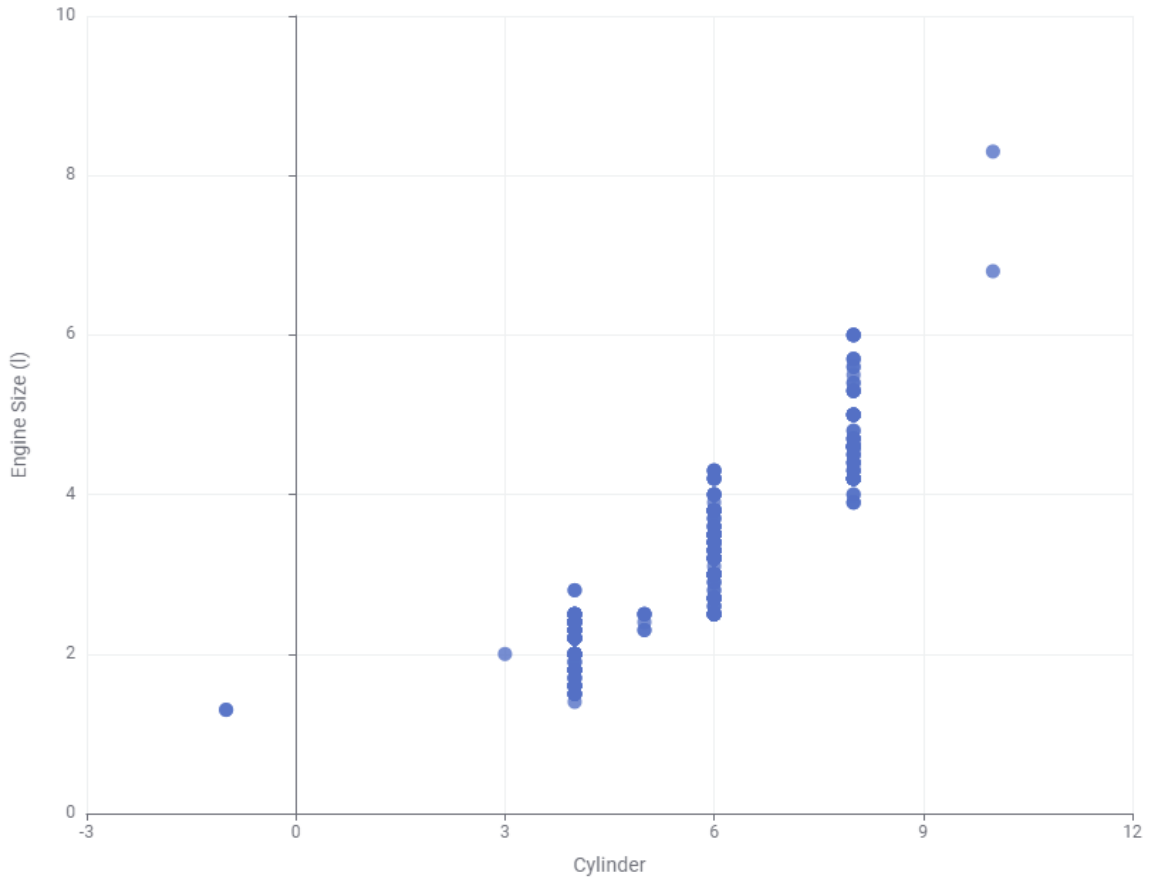


Figure 31: Correlation Analysis between Engine Size and Cyl

Figure 31 is the visual presentation of the non-visual correlation analysis between Engine Size (l) and Cyl. This reveals that there exists a High Positive Correlation coefficients (r) of 0.9 i.e. highly positively correlated.

4.3.4 Correlation Analysis between Cyl and HP

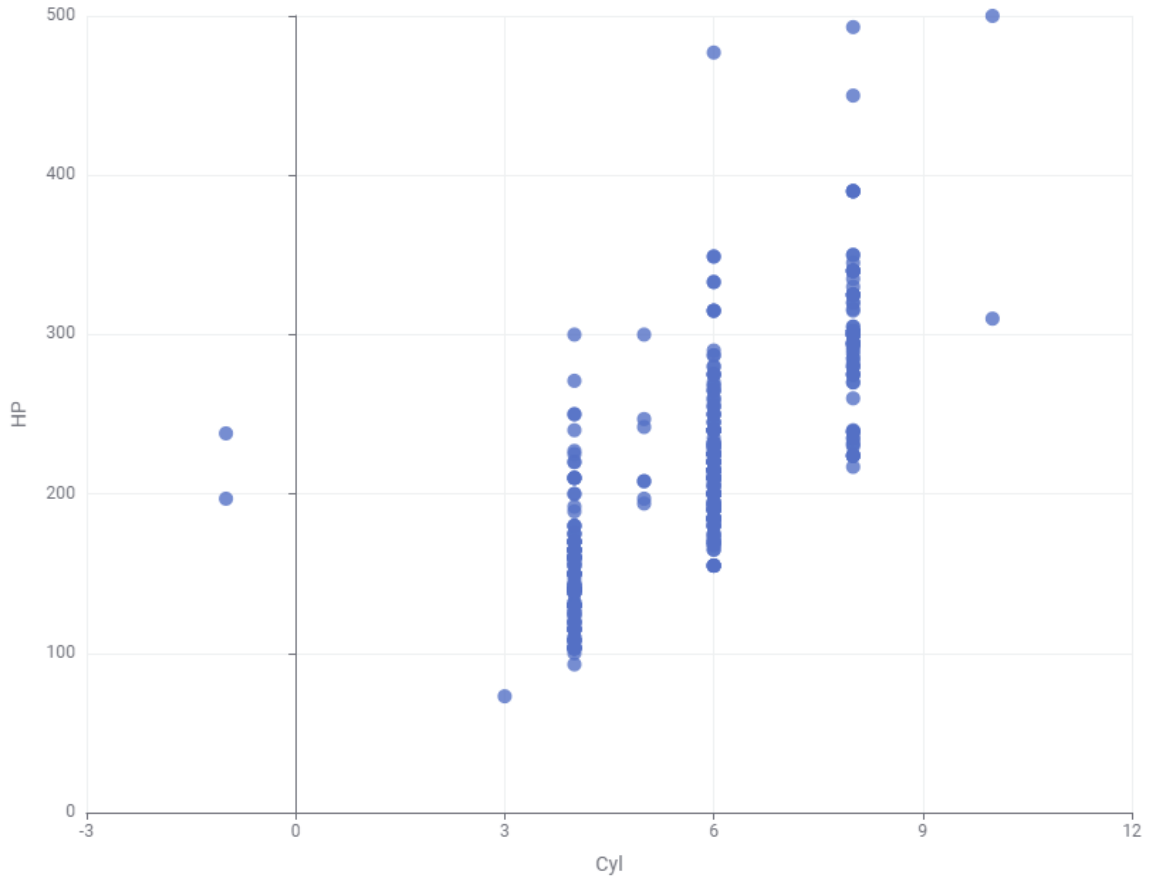


Figure 32: Correlation between Cyl and HP

Figure 32 is the visual presentation of the non-visual correlation analysis between Engine Size (l) and Cyl. This reveals that there exists a High Positive Correlation coefficient (r) of 0.8 i.e. highly positively correlated.

4.4 Visual representation and Explanation of some Grouping and Binning

The car dataset was grouped into different bins. Some numerical variable were converted into bins using the 'Auto-Binner' and 'Groupby' nodes in knime analytics. The numerical values that was grouped were Highway Miles Per Gallon (Hwy MPG), City MPG, Retail Price and Dealer Cost. The binning method used was 'fixed number of bins' which was set at 5 of equal frequency with 'border' bin name. This was set at the bin setting in the knime analytics auto-binner. Below are the visual representation of the binned variables and their correspond explanations.

4.4.1 Count of Vehicle by Retail Price (Binned)

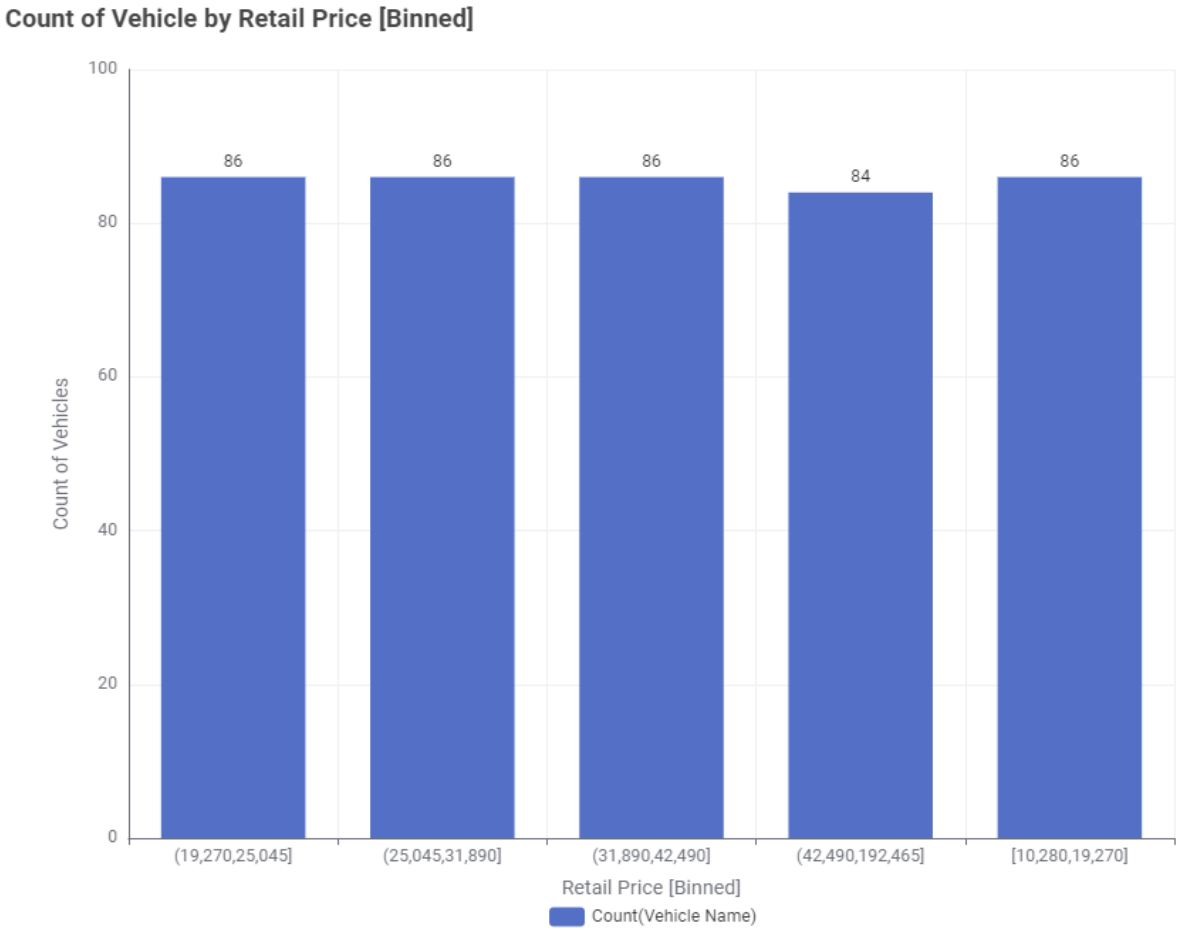


Figure 33: Count of Vehicle By Retail Price (Binned)

From figure 33, the bin represents the lower and the upper limits of each retail price bin. It shows the count of vehicles that falls into the different bins of retail price. With the exception of bin [10, 280, 19,270] which has 84 vehicles in the category, the rest bins have equal number of count of vehicles i.e. 86.

4.4.2 Count of Vehicle By Dealer Cost (Binned)

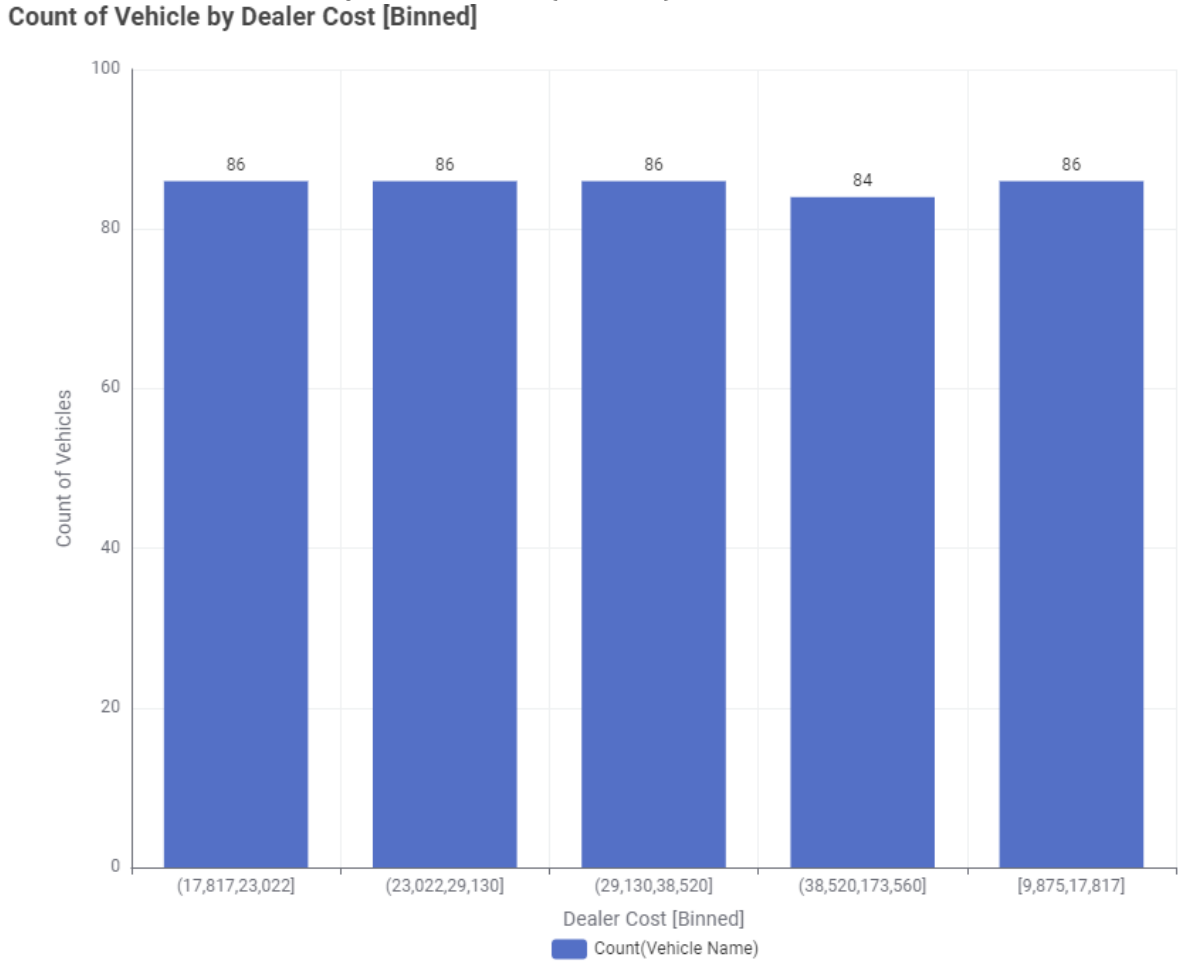


Figure 34: Count of Vehicle By Dealer Cost (Binned)

From figure 34, the bin represents the lower and the upper limits of each dealer cost bin, it shows the count of vehicles that falls into the different bins of dealer cost. Like the retail price, the bin [38, 520, 173, 560] has 84 vehicles in the category, the rest bins have equal number of count of vehicles i.e. 86 each.

4.4.3 Count of Vehicles By City MPG (Binned)

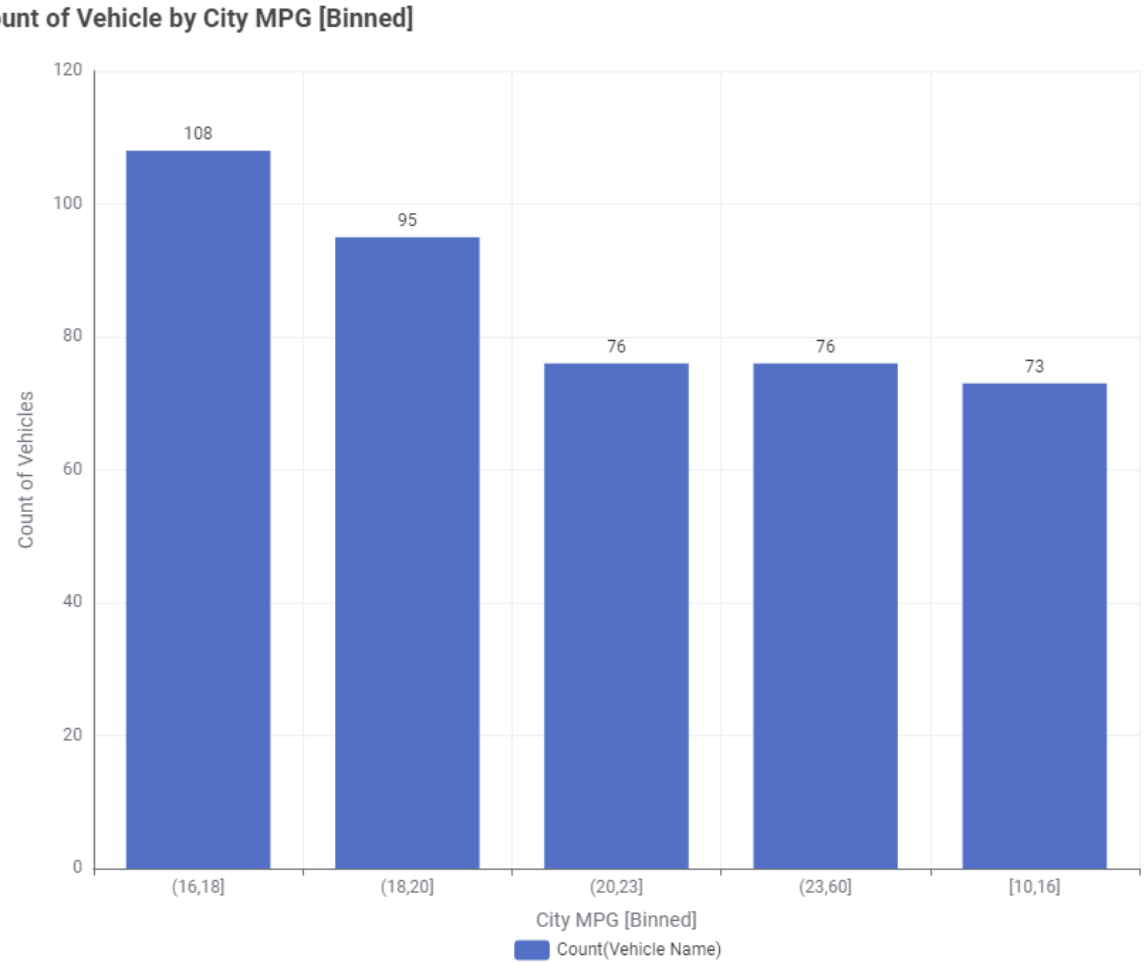


Figure 35:Count of Vehicle By City MPG (Binned)

In figure 35, the bin represents the lower and the upper limits of each City MPG bin, it reveals at a glance the count of vehicles in each city MPG bin. The bin (16, 18] has the highest count of 108 (no of vehicles in that group of fuel consumption withing the city) while the bin (10,16) has the lowest count of 73.

4.4.4 Count of Vehicles By Highway MPG (Binned)

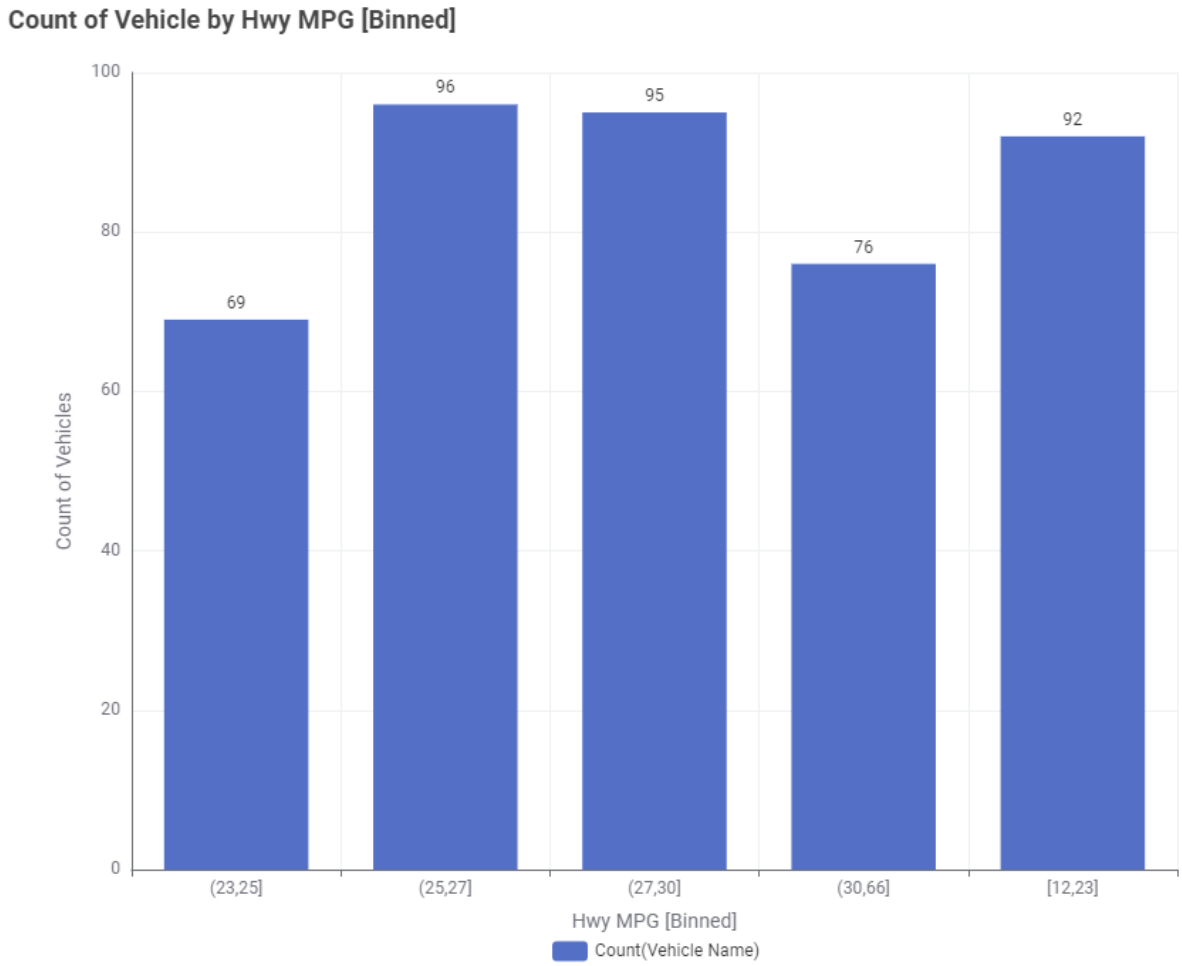


Figure 36:Count of Vehicle By Hwy MPG (Binned)

From figure 36, the bin represents the lower and the upper limits of each Hwy MPG bin, it can be seen that the bin (25,27) had the highest count of vehicles of 96 while the bin with the lowest count of vehicle is the (23,25) bin.

4.5 Visual Representation of Some Nominal Attributes

The car dataset lacked one of the conditions for this activity in that it does not contain any nominal attribute. This was overcome by converting some variables into nominal attributes. The variables that converted into nominal were Sedan, Sports Car, SUV, Wagon, Minivan, Pickup, AWD (All Wheel Driver), and RWD (Rear Wheel Driver). These were vehicles of different that different brands of car manufacturers produced according to the dataset.

4.5.1 Tasks

The following tasks were undertaken in knime analytics during the conversion of the listed variables to nominal attribute.

a. Lookup Table Creation: A lookup table was created by using the table creator node in the knime analytics. The converted variables two values i.e. 1 and 0. These values were replaced with Yes and No (1 = Yes, 0 = No) by creating a lookup table were with replacement values that the main table will reference to replace the 1 and 0 values. Below is section of the look up table. The table Creator node was used for this task.

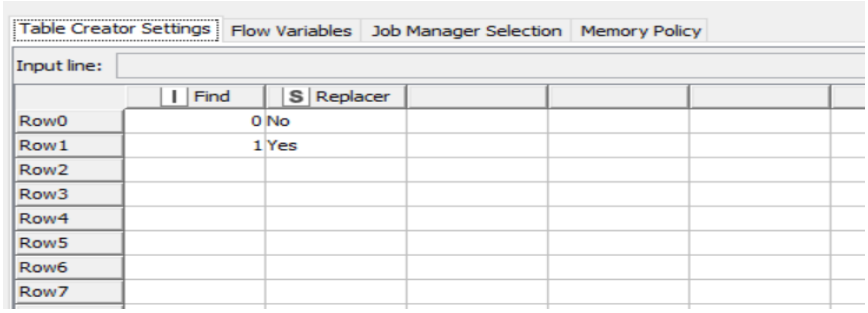


Figure 37: Pictorial Representation of the Lookup Table

b. Cell Replacement: Each of the cell containing the values 1 and 0 were replaced with yes and No respectively in each of the selected variable. The 'Cell Replacer' node was used to achieve this.

c. Data Grouping: the data grouped into two categories of Yes and No which were interpreted as:
   i. Yes: The vehicle type is available for that vehicle name
   ii. No: The vehicle type is not available in that vehicle name

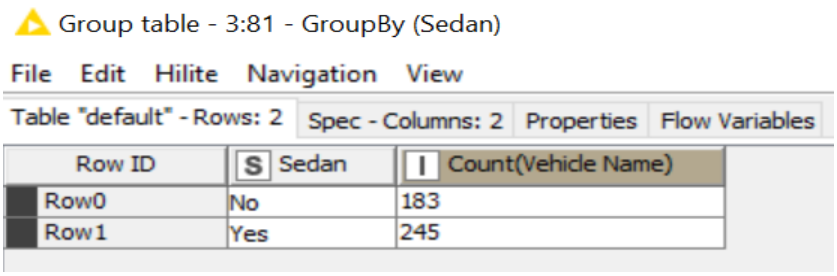


Figure 38: Pictorial representation of the GroupBy Output

d. Visualization of the Count of Vehicle name (Grouped Data by Vehicle Type): This was achieved by the use of Pie Chart (Labs) node.

Below are the Pie charts visualization of the count of vehicle grouped data by vehicle type.

4.5.1 Visualization of Count of Vehicles By Different Vehicle Types

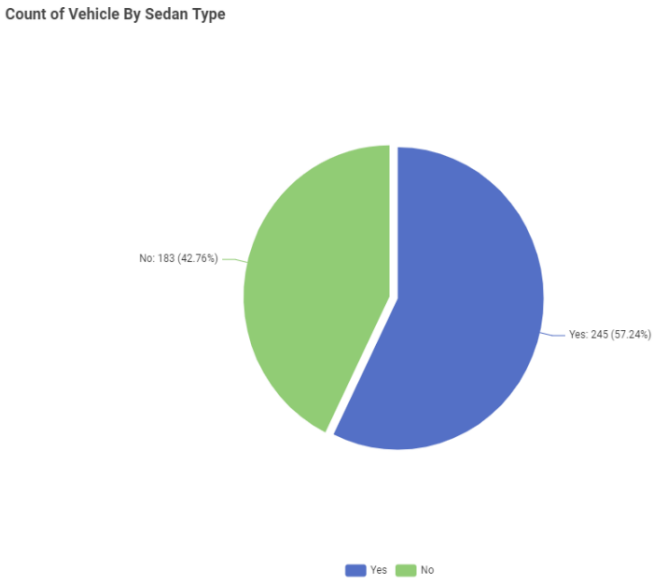


Figure 39 reveals the distribution of Sedan type of vehicle into Yes and No. It shows that there are 245 (57.24%) records available in the Sedan type while only 183 (42.76%) records does not have the Sedan type of vehicle.

Figure 39: Count of Vehicle By Sedan Type

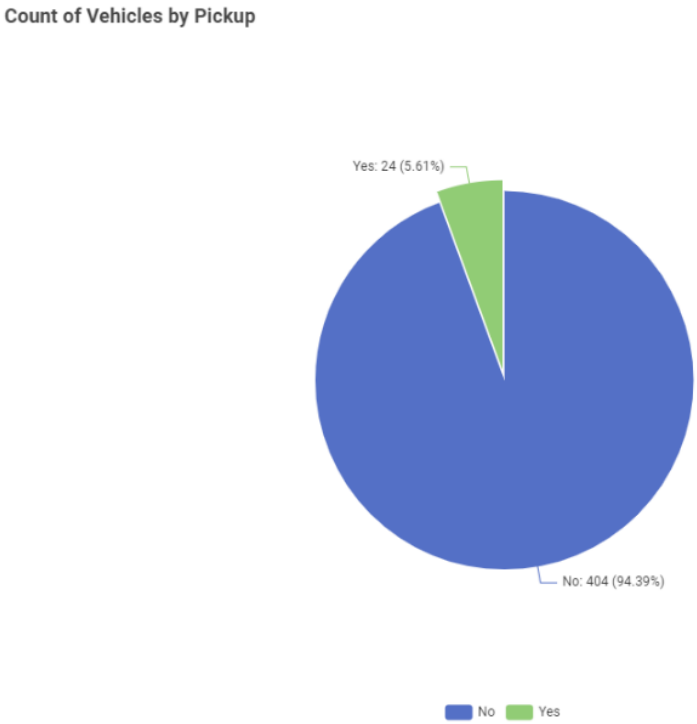


Figure 40 reveals the distribution of Pickup type of vehicle into Yes and No. It shows that there are 24 (5.61%) records available in the Pickup type while only 183 (42.76%) records does not have the Pickup type of vehicle.
vehicle.

Figure 40: Count of Vehicle By Pickup Type

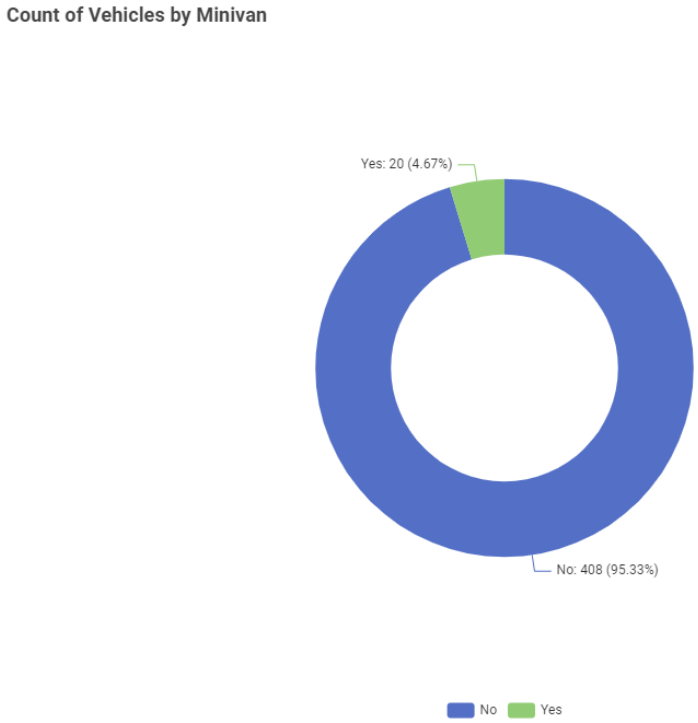


Figure 41 shows the distribution of the Minivan Type of vehicles. It reveals that there are 20 (4.47%) Minivan vehicle type distributed over different brands while the there were 408 (95.33%) which represents the count of brands which does not the Minivan type of vehicle.

Figure 41: Count of Vehicle By Minivan Type

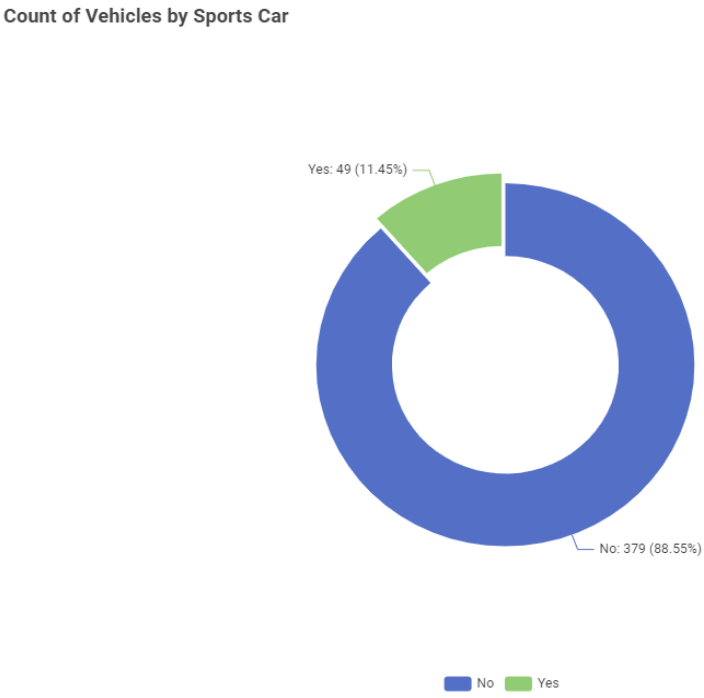


Figure 42 is the distribution of the Sports Car Type of vehicles. It reveals that there are 49 (11.45%) Sports Cars vehicle type distributed over different brands while the there were 379 (88.6%) which represents the count of brands which does not have the Sports car type of vehicle.

Figure 42: Count of Vehicle By Sports Car Type

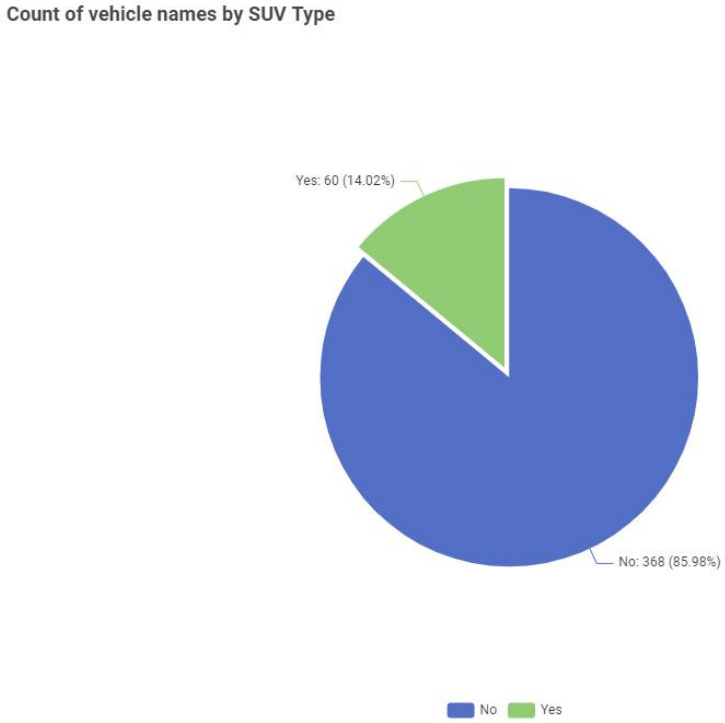


Figure 43 is the distribution of the SUV Car Type of vehicles. It reveals that there are 60 (14.02%) SUV Cars vehicle type distributed over different brands while the there were 368 (85.98%) which represents the count of brands which does not have the SUV car type of vehicle.

Figure 43: Count of Vehicle By SUV Car Type

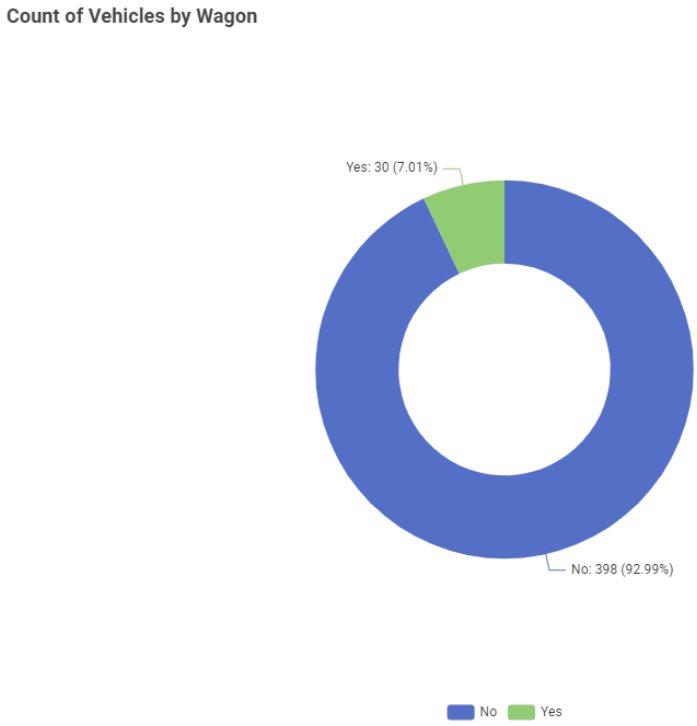


Figure 44 is the distribution of the Wagon Car Type of vehicles. It reveals that there are 30 (7.01%) Wagon Cars vehicle type distributed over different brands while the there were 398 (98.99%) which represents the count of brands which does not have the Wagon car type of vehicle.

Figure 44: Count of Vehicle By Wagon Type

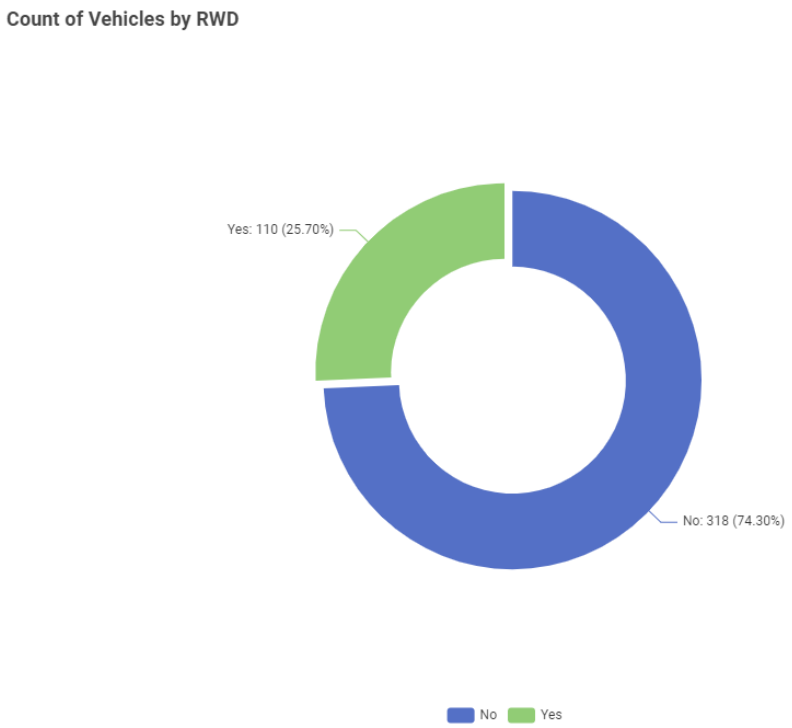


Figure 45 is the distribution of the RWD (Rear Wheel Driver) Type of vehicles. It reveals that there are 110 (25.70%) RWD Cars vehicle type distributed over different brands while the there were 318 (74.30%) which represents the count of brands which does not have the RWD type of vehicle.

Figure 45: Count of Vehicle By RWD Type

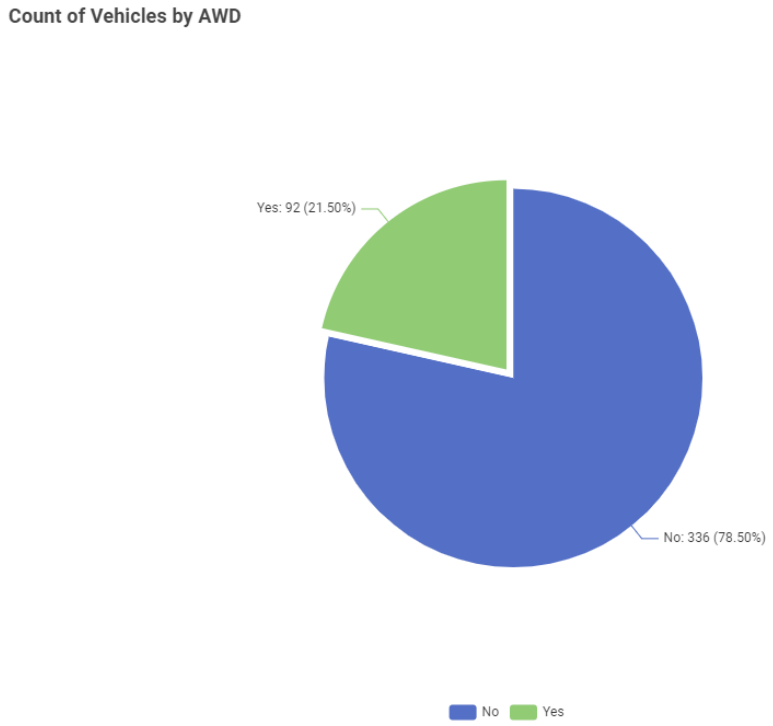


Figure 42 is the distribution of the AWD (All Wheel Drive) Type of vehicles. It reveals that there are 92 (21.50%) AWD Cars vehicle type distributed over different brands while the there were 336 (78.50%) which represents the count of brands which does not have the AWD car type of vehicle.

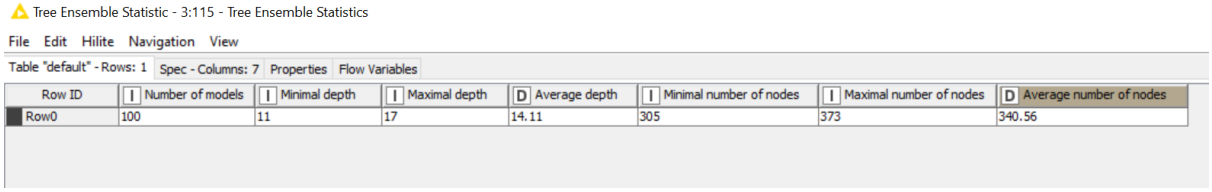
Figure 46: Count of Vehicle By AWD Type

4.6 Regression Analysis

Regression analysis is the statistical measure used to understand, amongst many variables, which is most important and relevant to have effect on the variable of interest. For example, in logistics and transportation industry, improving drivers' driving behavior and monitoring sharp practices are important issues driving business owners crazy, hence having a negative impact on the bottom line of the company i.e. Profit. The bottom line of any organizations to make and increase profit (this is the variable of interest – dependent variable) can be many factors , which includes, but are not limited, delay at customers site, drivers sharp practices, wear and tear on assets, fatigue on the wheel for not observing hours of service etc. Amy 2015 was of the opinion that regression analysis helps to sort out and identify which of these factors have the most significant impact. It helps in answering the questions: Which factors are of higher importance?, Which of these factor can be ignored?, What is the relationship between these factors?, and the most important question is the confirmation of the certainty of the most important factor.

The variable of interest in this case is the retail price. The regression analysis carried out was to identify which of the variables has the most significant effect on the prediction of the retail prices of each of the vehicles types in the dataset. Below is the decision tree from the regression analysis carried out using the knime Random Forest Learner and Random Forest Predictor.

Tree ensemble statistics node was used to calculate the summary statistics of the implemented random forest.


Tree Ensemble Statistic - 3:115 - Tree Ensemble Statistics

File Edit Hilite Navigation View

Table "default" - Rows: 1 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	Number of models	Minimal depth	Maximal depth	Average depth	Minimal number of nodes	Maximal number of nodes	Average number of nodes
Row0	100	11	17	14.11	305	373	340.56

Figure 47: Summary Statistics Table for Random Forest

From figure 47, it shows that the number of models was 100 with minimal depth of 11, maximal depth of 17, 305 minimal no of node and 373 maximal no of nodes.

Below is the output statistics from the Random Forest Learner.

Attribute Statistics - 3:49 - Random Forest Learner (Regression) (ML using Test Data)

File Edit Hilite Navigation View

Table "Tree Ensemble Column Statistic" - Rows: 6 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	#splits (level 0)	#splits (level 1)	#splits (level 2)	#candidates (level 0)	#candidates (level 1)	#candidates (level 2)
Dealer Cost	38	56	145	38	56	149
Engine Size (l)	12	22	60	32	77	137
Cyl	13	28	26	24	62	129
HP	28	54	83	37	71	133
City MPG	9	23	37	34	56	124
Hwy MPG	0	17	34	35	78	128

Figure 48: Attribute Statistics (Random Forest Learner)

From the above classification looking at the split (level 0), dealer cost has been considered for the top split criterion 38 out of 100 decision and then chosen many times for the top split criterion. On the contrary, input feature Hwy MPG has been considered 35 times as a candidate for the top split criterion but was never chosen as split criterion. This indicates that dealer cost isa more important feature for classification task than Hwy MPG i.e.. It is more statistically significant in prediction of the retail than the Hwy MPG.

Prediction output - 3:51 - Random Forest Predictor (Regression) (Prediction of the)

File Edit Hilite Navigation View

Table "default" - Rows: 299 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Retail P...	Dealer ...	Engine ...	Cyl	HP	City MPG	Hwy MPG	Retail P...	Retail P...	model c...	Prediction (Retail Price)	Prediction (Retail P...
Row0	43,755	39,014	3.5	6	225	18	24	41,029.324	8,288,250.938	35	42,803.438	4,558,972.242
Row2	36,945	33,337	3.5	6	265	17	23	34,604.221	12,893,005....	36	36,092.4	5,826,939.183
Row3	89,765	79,978	3.2	6	290	17	24	73,073.833	940,606,16...	30	84,533.675	335,319,090.967
Row4	23,820	21,761	2	4	200	24	31	23,591.095	4,314,669.192	37	23,713.805	1,650,704.273
Row5	33,195	30,299	3.2	6	270	20	28	32,794.275	5,974,639.89	34	33,058.753	2,027,944.972
Row6	26,990	24,647	2.4	4	200	22	29	25,877.896	8,570,612.336	39	26,605.039	3,897,123.109
Row7	25,940	23,508	1.8	4	170	22	31	24,884.418	1,659,583.523	39	25,366.061	879,617.143
Row8	31,840	28,846	3	6	220	20	28	32,385.631	4,517,173.569	33	32,045.608	1,590,625.096
Row10	34,480	31,388	3	6	220	18	25	34,922.938	9,459,677.105	36	34,816.94	4,456,089.84
Row11	33,430	30,366	3	6	220	17	26	35,316.155	87,394,253....	39	34,329.721	36,139,816.199
Row12	44,240	40,075	3	6	220	18	25	40,853.39	7,948,498.172	44	42,318.37	6,082,930.857
Row13	35,940	32,506	1.8	4	170	23	30	32,899.362	52,846,456....	29	34,959.778	17,722,354.115
Row15	36,640	33,129	3	6	220	20	27	35,497.688	4,315,650.445	40	36,187.108	2,021,788.615
Row17	39,640	35,992	3	6	220	18	25	40,249.375	6,510,319.934	40	39,948.812	3,704,793.528
Row19	69,190	64,740	4.2	8	330	17	24	72,085.095	87,349,005....	32	70,219.308	29,890,698.045
Row20	84,600	76,417	4.2	8	450	15	22	108,409.206	1,821,774,0...	42	94,578.158	894,421,329.83
Row22	48,040	43,556	4.2	8	340	14	20	51,555.422	130,627,70...	35	49,267.258	47,710,719.247
Row23	35,940	32,512	1.8	4	180	20	28	34,081.391	13,547,638....	33	35,309.794	5,231,654.349
Row24	37,390	33,891	1.8	4	225	20	28	36,129.327	9,555,881.001	28	37,006.417	2,937,948.41
Row25	40,590	36,739	3.2	6	250	21	29	40,344.417	14,912,768....	36	40,502.523	5,295,159.504
Row27	37,995	34,800	2.5	6	184	19	27	37,382.222	5,344,754.873	39	37,768.925	2,197,608.043
Row28	28,495	26,155	2.5	6	184	20	29	28,946.06	24,875,247.14	36	28,646.517	8,896,238.399
Row34	37,245	34,115	3	6	225	20	29	38,038.25	8,447,793.013	40	37,544.75	3,816,865.671
Row35	39,995	36,620	2.5	6	184	19	28	38,347.639	5,000,990.694	36	39,355.133	2,447,708.68
Row36	44,995	41,170	3	6	225	20	30	41,897.265	14,701,891....	34	43,941.77	7,075,715.815
Row37	54,995	50,270	4.4	8	325	18	26	56,663.423	34,651,441....	37	55,808.602	16,879,876.782
Row38	69,195	63,190	4.4	8	325	18	26	66,020.926	51,118,849....	36	68,034.31	21,275,120.673
Row42	37,000	33,873	3	6	225	16	23	35,485.923	21,187,986....	41	36,357.208	9,119,081.363
Row43	52,195	47,720	4.4	8	325	16	22	56,051.542	106,312,34...	40	53,737.617	45,486,174.93
Row45	41,045	37,575	3	6	225	21	29	40,260.586	8,063,931.419	35	40,741.778	2,938,519.36
Row47	26,470	24,282	3.8	6	205	20	29	27,366.652	11,950,498....	35	27,006.491	5,652,456.619
Row48	32,245	29,566	3.8	6	205	20	29	32,419.651	5,646,090.947	43	32,385.202	2,758,179.212
Row49	35,545	32,244	3.8	6	205	20	29	34,247.895	8,210,500.853	38	34,762.189	4,157,031.989
Row50	40,720	36,927	3.8	6	240	18	28	39,483.152	77,928,552....	33	40,157.152	26,635,797.202
Row52	28,345	26,047	3.8	6	240	18	28	28,113.488	13,029,870....	33	28,337.204	4,735,344.616
Row53	24,895	22,835	3.8	6	200	20	30	24,930.873	2,352,261.512	42	24,869.434	1,009,321.472
Row54	26,545	24,085	3.4	6	185	19	26	25,680.823	2,956,892.367	38	26,129.093	1,470,251.156
Row56	45,445	41,650	4.6	8	275	18	26	47,701.274	27,464,866....	37	46,717.095	12,901,258.323
Row57	50,595	46,362	4.6	8	300	18	26	51,007.874	96,932,886....	42	50,717.17	40,343,063.275

Figure 49: Prediction Output (Random Forest Predictor)

There is a statistically significance reduction in the predicted retail prices of the each of the products. In some few cases, it increased but not significantly. Below is a snapshot of the output table from the prediction of the retail price.

As previously explained in figure 47, there was a decision tree resulting from the prediction analysis which explains the importance of each of the attributes on the retail price. Below is the picture of the snapshot of the decision tree.

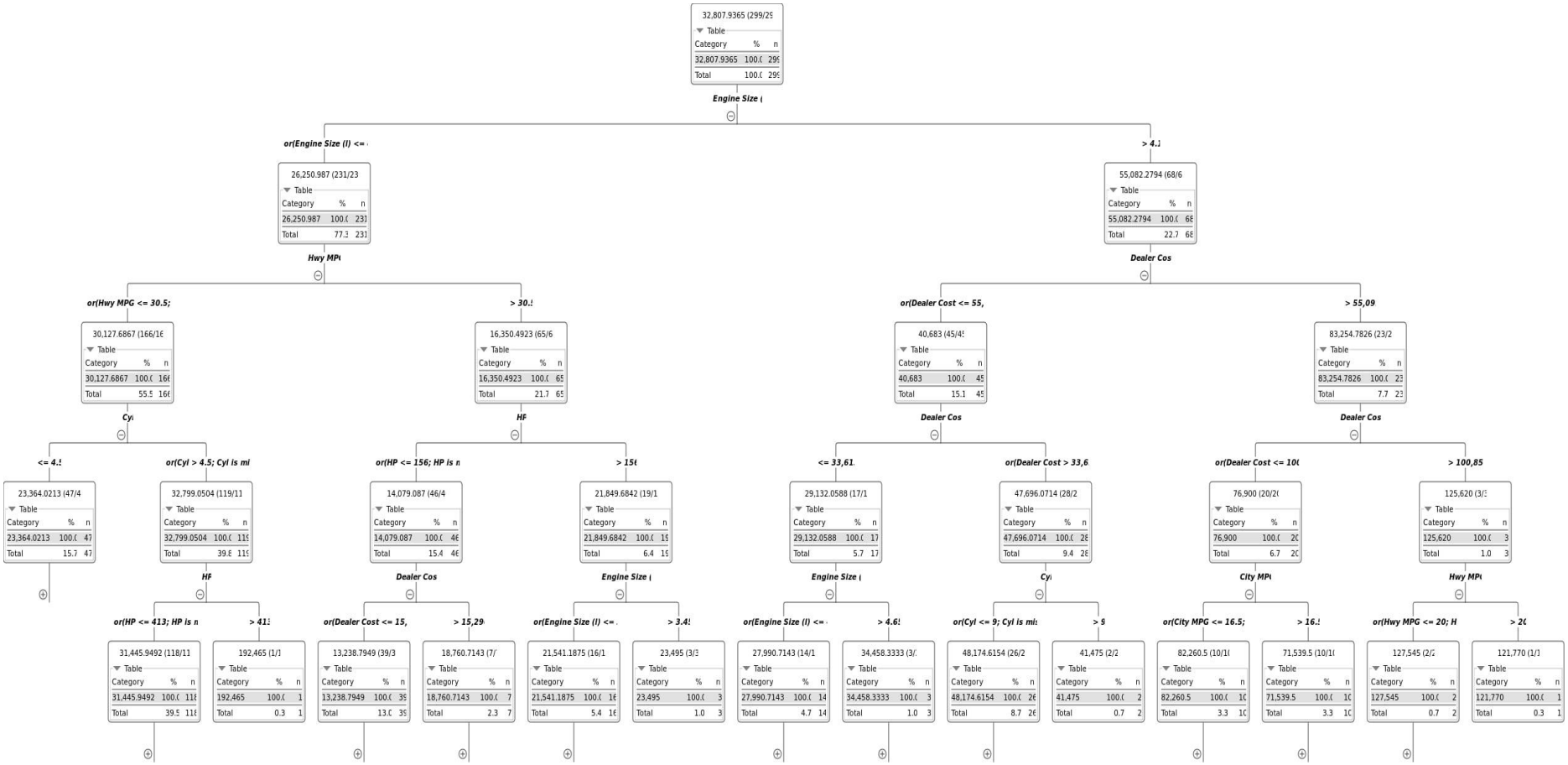


Figure 50: A Snapshot of the Decision Tree.

5.0 Workflow Design for Visual Representation of Tabular Data Using Knime Analytics

The tabular dataset of choice was the MFG10YearTerminationData.csv which was found on Kaggle website at (https://www.kaggle.com/datasets/HRAnalyticRepository/employee-attrition-data). It was a publicly available Fake termination data which can be used to try to predict employee attrition.

This section aims at designing a Knime workflow to visualize the data in appropriate visual representations. Below is the pictorial representation of the Knime workflow for the visual presentations.

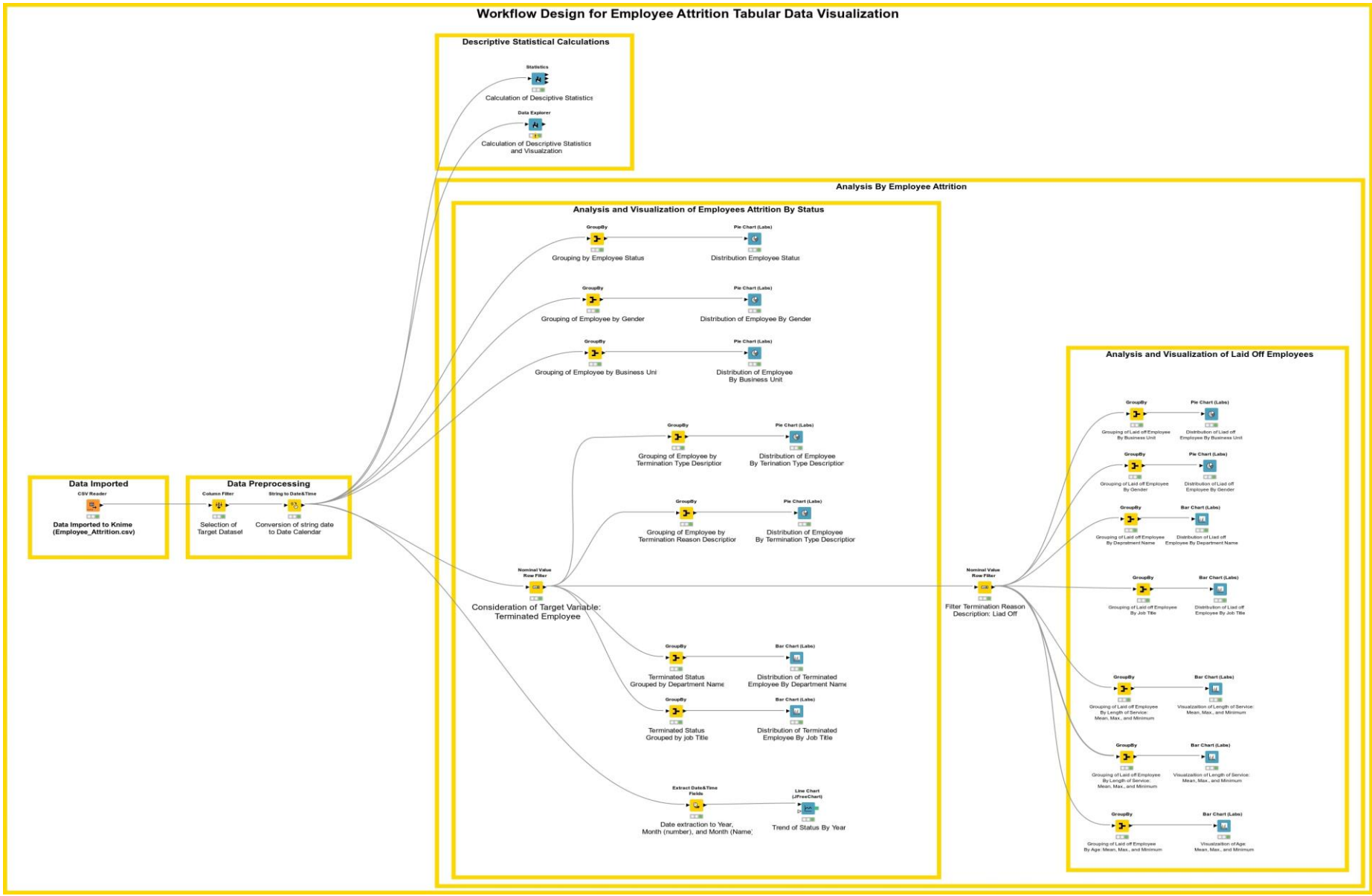


Figure 51: Designed Knime Workflow for Visualization of Employee Attrition Dataset

5.1 Output and Findings of the Knime Workflow Design for Tabular Dataset Visualization

Several tasks were carried out while designing the knime visualization workflow, they are:

a. Data Importation: The employee attrition tabular datasets was imported using the knime csv reader node.
b. Data Preprocessing: Some preprocessing were done using two knime nodes

    i. Selection of Target Dataset: This was done to remove the columns which are not relevant to the analysis of the Human Resources data in analysing employee attrition. The node used was column filter node.
    ii. Conversion of String Date (String) to Date (Calendar): The terminationkey_date column contains the date which each of the employees whose appointment were terminated by the employer. The column was stored and imported in String data type which was not the appropriate data type for knime for analysis. This was fixed by the use of string to date and time node.

c. Descriptive Statistical Calculations: Some descriptive statistics were calculated using the statistics and data explorer nodes. The data explorer, in addition to calculating some statistics, added the visuals of each of the statistical measures.
d. Analysis and Visualization of Employees Attrition By Status: The aim of choosing this data is to investigate the employee turnover in the company and possibly to investigate the reasons behind their attrition. The analysis was done in the following order: Employee Attrition By Status (focusing on the reason for Termination), Employee Attrition By Termination Reason Description (with emphasis on Laid Off Category). Each of these category were further analyzed and visualized using different but appropriate node for analysis and visualization.

5.1.1 Descriptive Statistics

Some descriptive statistics were carried out using the statistics and data explorer nodes. Focusing on the output of the data explorer node, some statistical measures were discovered with their corresponding visuals. The output gave numeric, nominal and general data previews.

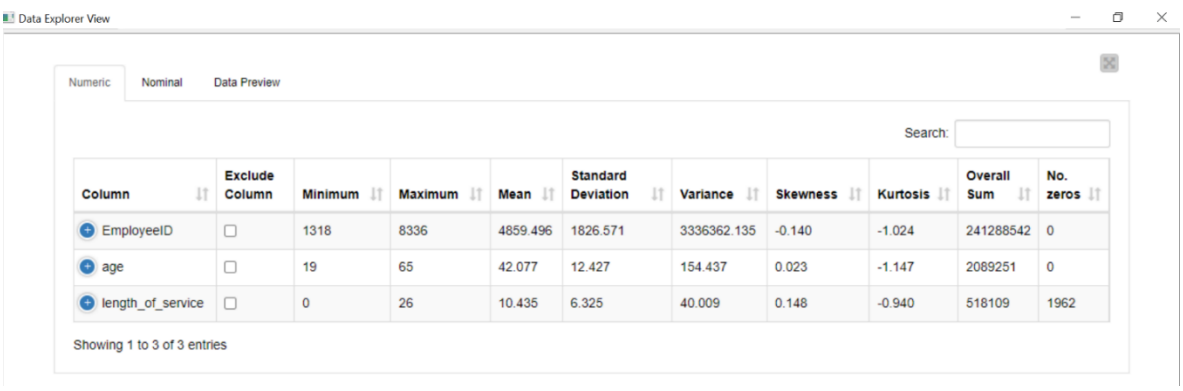


Figure 52: Snapshot of the Data Explorer Statistical Measures.

a. The numeric view of the data explorer helps to visualize the statistical measures of these numerical data: Employee ID, Age, and Length of service, however the last two are of immense consideration.

i. Age

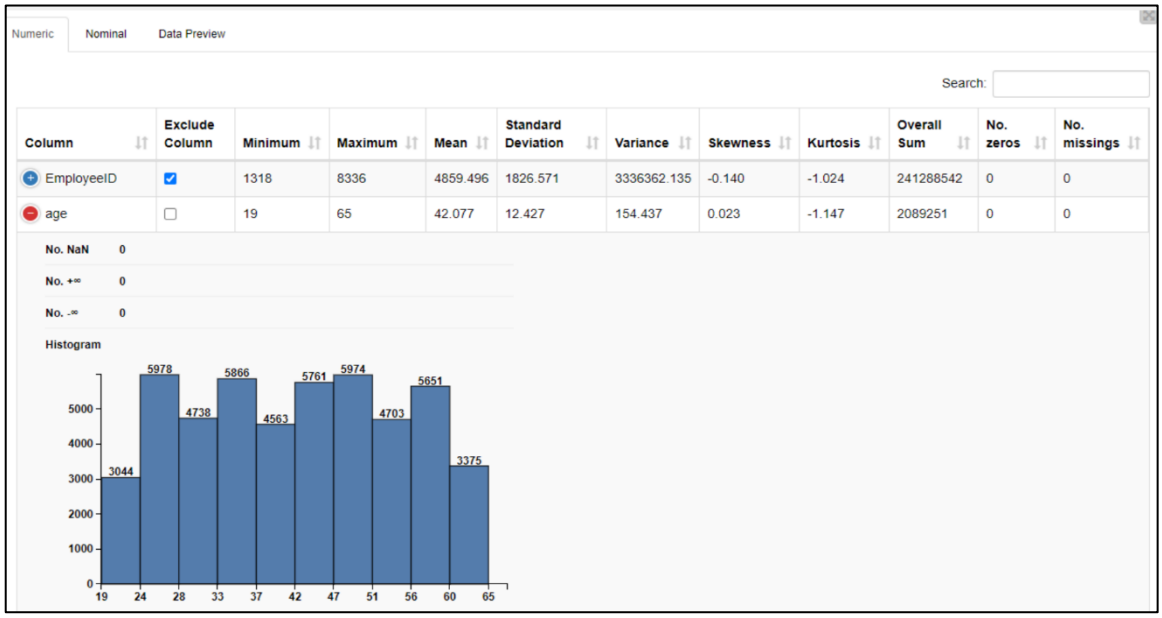


Figure 53: Visualization of Employee Statistical Measures by Age

From figure 53, it showed the frequency of the age of the employee. Employee between the ages 24 to 28 has the highest count of 5, 978 while the employee between the age range 19 to 24 has the lowest number of the employee. This reveals that the majority of the companies employee falls between the age range 24 to 28.

ii. Length of Service

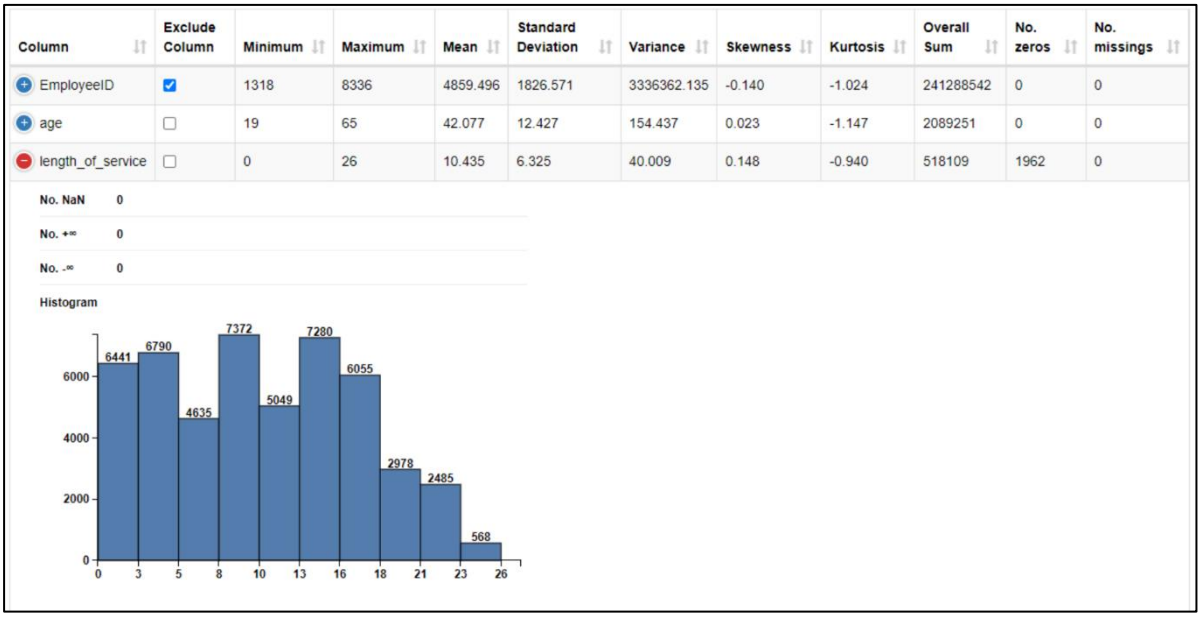


Figure 54: Visualization of Employee Statistical Measures by Length of Service

Figure 54 reveal that majority of the company's employees falls between 8 to 10 years of service i.e. 7,372 in number while a fewer number (568)falls between 25 to 26 years of length of service. This means that a fewer category of the company's employees were old people who were closer to their retirement age among the active employee.

b. The nominal view gave outputs on the statistical measures of the 8 nominal data attributes and their corresponding visuals.

Figure 55: Snapshot of the Data Explorer Nominal Attributes

5.1.2 Findings from the Visualization of the Tabular Data

a. Distribution of Employees by Employment Status

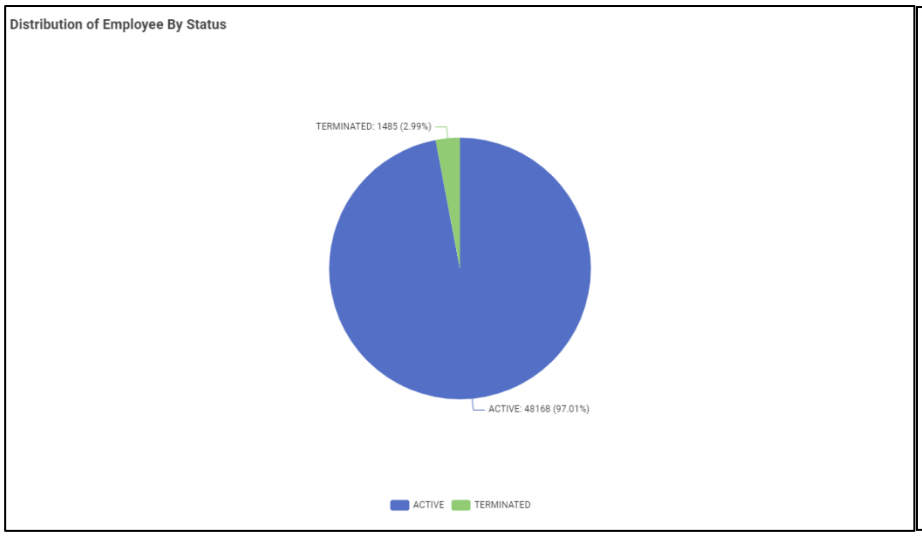


Figure 56 is the distribution of employee by status. This reveals that out of the total 49,653 employees, about 97% (48,168) were still active while the remaining 3% (1, 485 have been terminated for one reason or the other.

Figure 56: Distribution of Employee By Status

b. Distribution of Employees By Gender

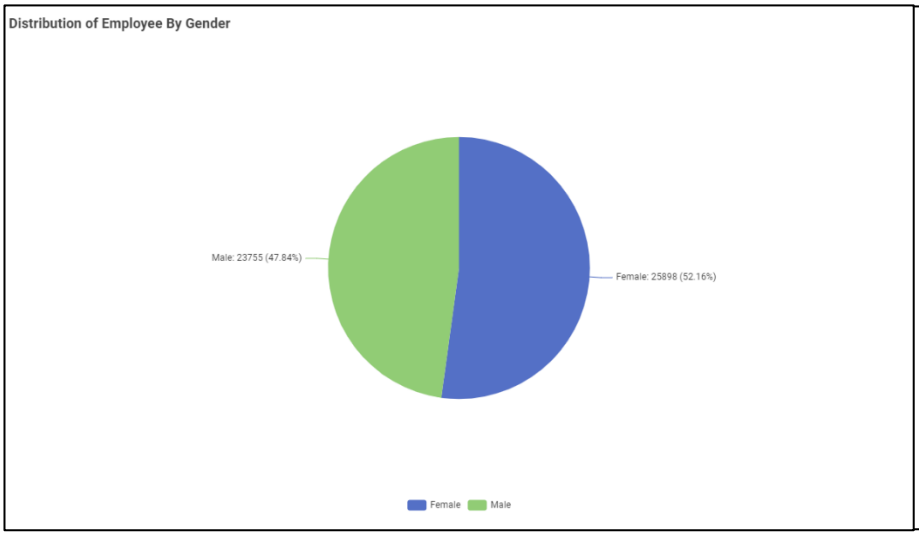
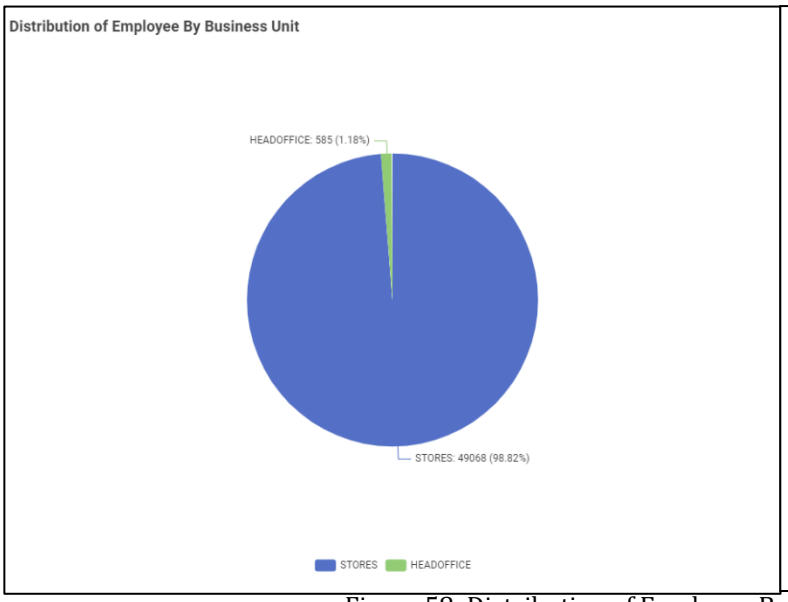


Figure 57 is the distribution of employee by Gender. This shows that out of the total 49,653 employees, about 52% (25, 898) were female while about 48% (23, 755) were male.

Figure 57: Distribution of Employee By Gender

c. Distribution of Employees By Business Unit



From figure 58, it can be seen that the company has only two major categorization of employee by office location i.e. Head Office and Stores. The Stores has the larger percentage of the employees with about 99% (49,068 employees) while the rest (about 1% i.e. 585 employees) are located at the Head office.

Figure 58: Distribution of Employee By Business Unit

From the above analysis, it was showed that the company has a total of 49, 653 employees. These were distributed by gender into 52% (25, 898) female and about 48% (23, 755) male. Of the total employees, distribution by business unit revealed 99% (49,068 employees) to be stores employees while the rest (about 1% i.e. 585 employees) were Head office employees. Considering the employment status, it was discovered that 97% (48,168) were still active while the remaining 3% (1, 485) have been terminated for one reason or the other. Based on the last statement, further analyses were carried out on the termination type description.

d. Terminated Employees By Termination Type Description

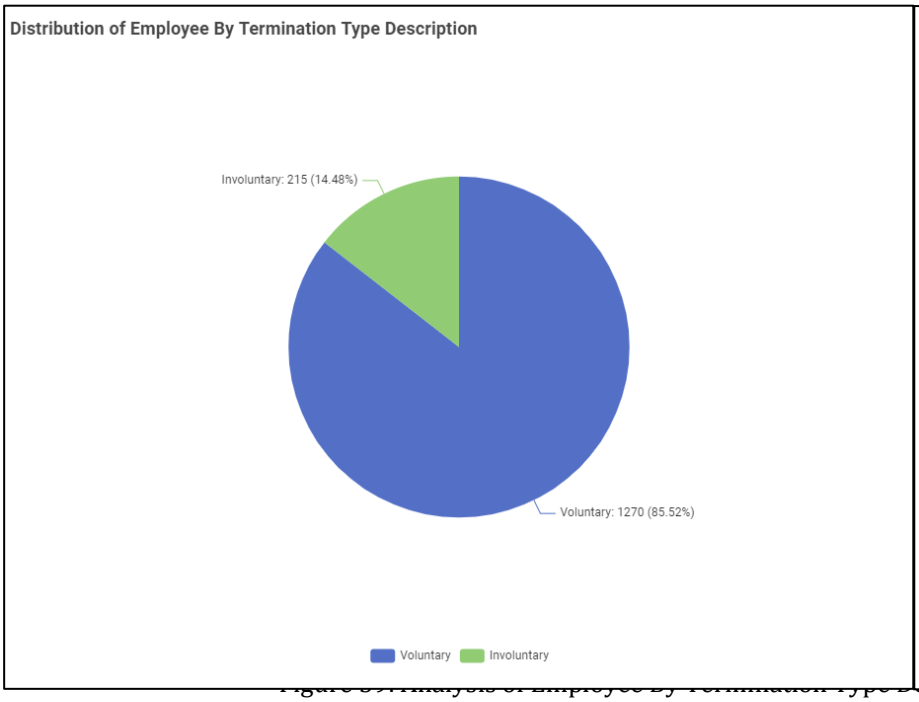
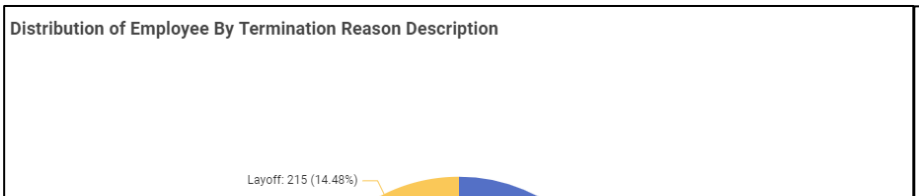


Figure 59 revealed that there were two mode of disengagement i.e. Voluntary and Involuntary. Of the 3% (1,485) that were terminated (figure 56), about 85% (1,270) disengaged voluntarily while approximately 15% (215 employees) were disengaged involuntarily.

Figure 59: Analysis of Employee By Termination Type Description

e. Terminated Employees By Termination Reason Description.

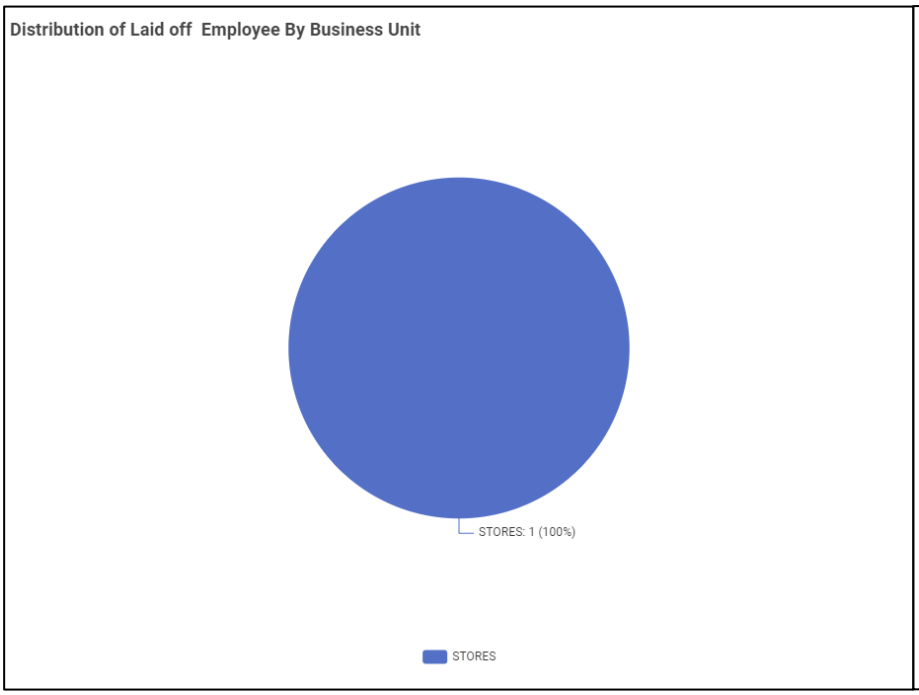


Further analysis of the 3% (1,485) that were terminated (figure 56) due to one reason or the other, about 85% (1,270) disengaged

Figure 60: Distribution of Terminated Employee By Termination Reason Description.

Based on the revelation in figure 60, the laid off employees were further concentrated on for further analysis to see the distribution into business unit , department name, job title, age, length of service and city name.

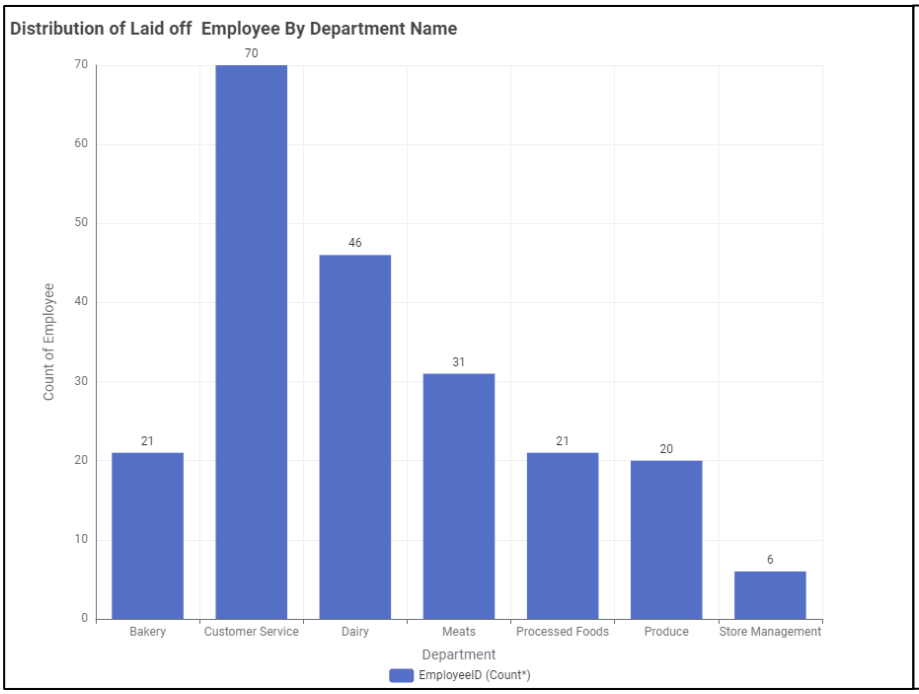
f. Laid Off Employees By Business Unit



From figure 61, it showed that all the 3% (1,485) that were terminated (figure 56) due to one reason or the other belong to only the Stores Unit. This follows that none of them belong to the Head Office.

Figure 61: Distribution of Laid Off Employees By Business Unit

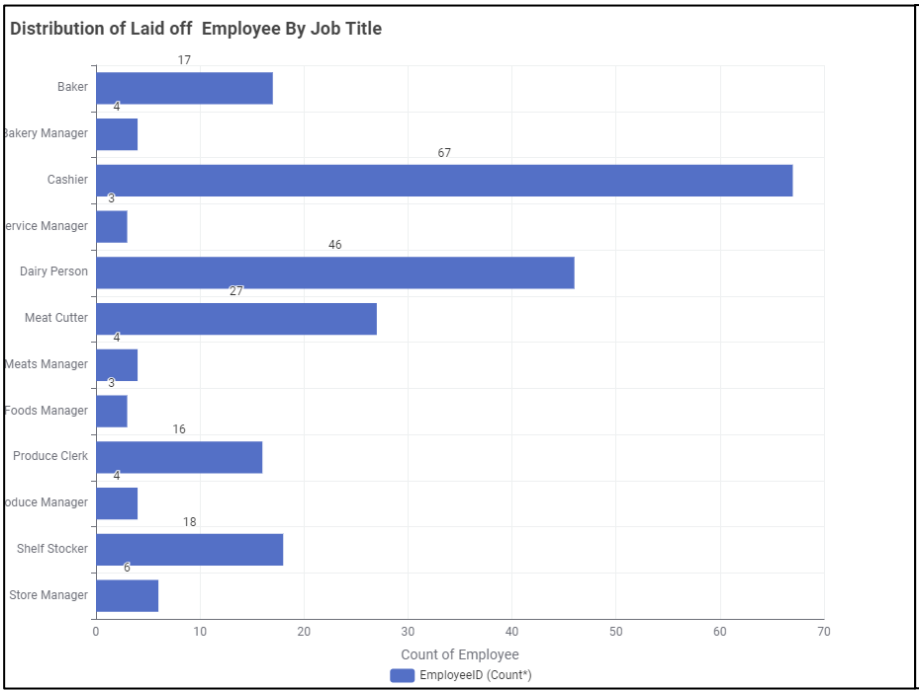
g. Laid Off Employees By Department Name.



From figure 62, further analysis of the 215 employees laid off by departmental basis revealed that about 33% (70) were in Customer Service (largest share of the 215 employees) with Dairy having a second largest share of 21.4% (46) while Store Management had the least share of the distribution i.e. about 3% (6 employees)

Figure 62: Distribution of Laid Off Employees By Department Name

h. Laid Off Employees By Job Title



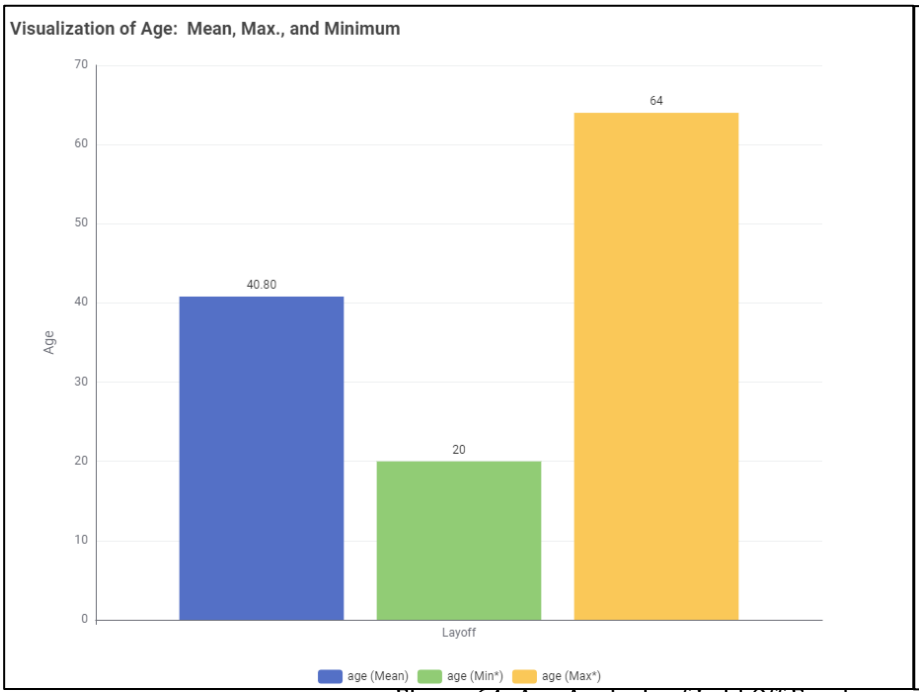
From figure 63, further analysis of the 215 employees laid off by job title basis revealed that about 31% (67) were in Cashiers (largest share of the 215 employees) with Dairy Person having a second largest share of 21.4% (46) while Service Manager and Foods Manager had the least shares of the distribution i.e. about 1.5% (3 employees) each respectively.

Figure 63: Distribution of Laid Off Employees By Job Title

Based on the discovery from figure 63 which showed that the cashiers have the lion share of the laid off employees, this may follow that because it was a department that deals with money (if the company was not using contactless form of payment), were engaging in sharp practices which may include but not limited to: lack of due diligence that could lead to loss of revenue, lack

of record keeping, etc. These are possible issues that may lead to termination (a non-pardonable offense) instead of possible warning for first offender.

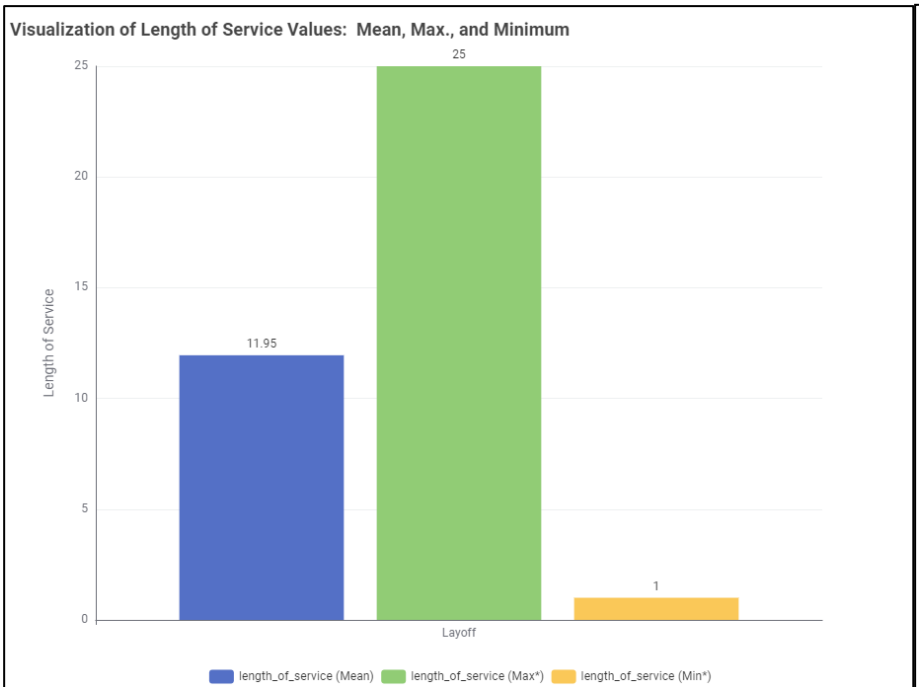
i. Laid Off Employees By Age



From figure 64, the maximum, mean and the minimum age of the category of the laid off employees were 64, 40.80 and 20 years respectively. Considering that cashiers have the largest share of the paid off staff and the mean age being approximately 41 years old, this may mean that because of the mid old age and possible on the job years of experience, complacency could be one of the reasons for large lay off among the employees of that mean age.

Figure 64: Age Analysis of Laid Off Employees

j. Laid Off Employees By Length of Service



From figure 65, the maximum, mean and the minimum length of service of the laid off employees were 25, 12 and 1 years respectively. Considering that cashiers have the largest share of the laid off staff and the mean age being approximately 12 years of service, this may also mean that majority of these employees were old enough in the company's system to know the policy guiding against unpardonable errors.

Figure 65: Laid Off Employees Length of Service

Conclusively, the employee attrition may not have a logical reason for employees being laid off however, with the revelation from the last 3 figures (i.e. figures 63 to 65), it showed that

sensitivity of the job description as designed by the job title, age and length of service may have direct impact (direct correlation) on the attrition.

5.2 Workflow Design for Visual Representation of Network or Geographical Data Using Knime Analytics (Activity 5.2).

This task involves the designing of knime workflow for the visualization of Network or geographical data, Based on data availability, a geospatial comma separated extension file (CitiesExt.csv) data about the population and temperature of different cities was downloaded from https://web.stanford.edu/class/cs102/datasets.htm . It was a concatenation of two datasets i.e. Cities.csv and Countries.csv.

5.2.1 Tasks

The data contained 8 columns and 213 rows. The column names were: city, country, population, EU, coastline, latitude, longitude, and temperature.

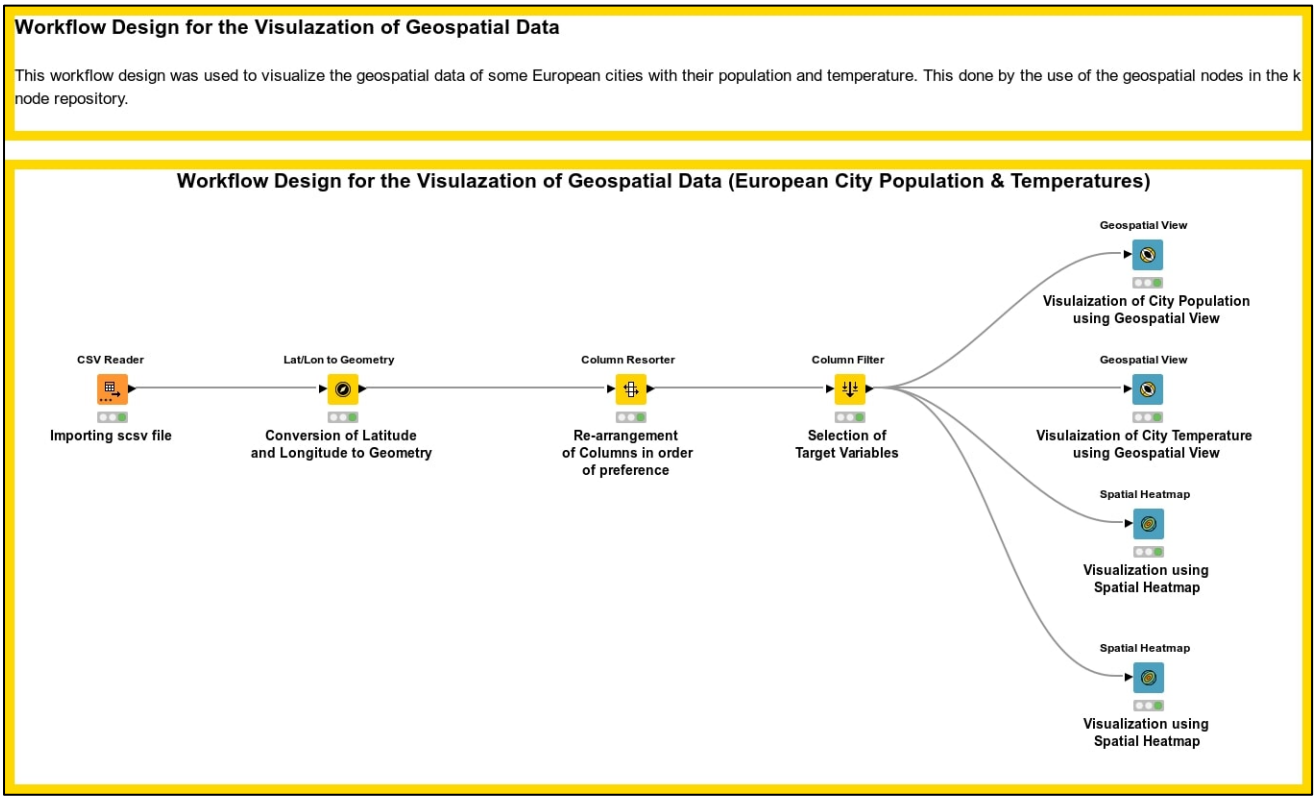


Figure 66: Workflow Design for Visualization of European City's Population and Temperature.

Several tasks were carried out in the workflow design.

a. Data Importation into Knime: The dataset was imported into knime using the csv reader node.

Row ID	S city	S country	D population	S EU	S coastline	D latitude	D longitude	D temperature
Row0	Aalborg	Denmark	5.69	yes	yes	57.03	9.92	7.52
Row1	Aberdeen	United Kingdom	65.11	yes	yes	57.17	-2.08	8.1
Row2	Abisko	Sweden	9.85	yes	yes	63.35	18.83	0.2
Row3	Adana	Turkey	79.62	no	yes	36.99	35.32	18.67
Row4	Albacete	Spain	46.06	yes	yes	39	-1.87	12.62
Row5	Algeciras	Spain	46.06	yes	yes	36.13	-5.47	17.38
Row6	Amiens	France	64.67	yes	yes	49.9	2.3	10.17
Row7	Amsterdam	Netherlands	16.98	yes	yes	52.35	4.92	8.93
Row8	Ancona	Italy	59.8	yes	yes	43.6	13.5	13.52
Row9	Andorra	Andorra	0.07	no	no	42.5	1.52	9.6
Row10	Angers	France	64.67	yes	yes	47.48	-0.53	10.98
Row11	Ankara	Turkey	79.62	no	yes	39.93	32.86	9.86
Row12	Antalya	Turkey	79.62	no	yes	36.89	30.7	11.88
Row13	Arad	Romania	19.37	yes	yes	46.17	21.32	9.32
Row14	Athens	Greece	10.92	yes	yes	37.98	23.73	17.41
Row15	Augsburg	Germany	80.68	yes	yes	48.35	10.9	4.54
Row16	Bacau	Romania	19.37	yes	yes	46.58	26.92	7.51
Row17	Badajoz	Spain	46.06	yes	yes	38.88	-6.97	15.61
Row18	Baia Mare	Romania	19.37	yes	yes	47.66	23.58	8.87
Row19	Balti	Moldova	4.06	no	no	47.76	27.91	8.23
Row20	Barcelona	Spain	46.06	yes	yes	41.38	2.18	15.78
Row21	Bari	Italy	59.8	yes	yes	41.11	16.87	15.15
Row22	Basel	Switzerland	8.38	no	no	47.58	7.59	6.68
Row23	Batman	Turkey	79.62	no	yes	37.89	41.14	14.16
Row24	Belfast	United Kingdom	65.11	yes	yes	54.6	-5.96	8.48
Row25	Belgrade	Serbia	8.81	no	no	44.82	20.47	9.85
Row26	Bergamo	Italy	59.8	yes	yes	45.7	9.67	9.12
Row27	Bergen	Norway	5.27	no	yes	60.39	5.32	1.75

Figure 67: A Snapshot of Original Geospatial Data

b. Geometry: Visualization of geospatial data on open street map (OSM) requires the dataset to contain a geometry column. This was absent in the original dataset downloaded however, it contained longitude and latitude columns for each city. These columns were converted to geometry by the use of Latitude/Longitude to Geometry node.

Table "default" - Rows: 213 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	city	country	population	EU	coastline	latitude	longitude	temperature	geometry
Row0	Aalborg	Denmark	5.69	yes	yes	57.03	9.92	7.52	POINT - EPSG:4326
Row1	Aberdeen	United Kingdom	65.11	yes	yes	57.17	-2.08	8.1	POINT - EPSG:4326
Row2	Abisko	Sweden	9.85	yes	yes	63.35	18.83	0.2	POINT - EPSG:4326
Row3	Adana	Turkey	79.62	no	yes	36.99	35.32	18.67	POINT - EPSG:4326
Row4	Albacete	Spain	46.06	yes	yes	39	-1.87	12.62	POINT - EPSG:4326
Row5	Algeciras	Spain	46.06	yes	yes	36.13	-5.47	17.38	POINT - EPSG:4326
Row6	Amiens	France	64.67	yes	yes	49.9	2.3	10.17	POINT - EPSG:4326
Row7	Amsterdam	Netherlands	16.98	yes	yes	52.35	4.92	8.93	POINT - EPSG:4326
Row8	Ancona	Italy	59.8	yes	yes	43.6	13.5	13.52	POINT - EPSG:4326
Row9	Andorra	Andorra	0.07	no	no	42.5	1.52	9.6	POINT - EPSG:4326
Row10	Angers	France	64.67	yes	yes	47.48	-0.53	10.98	POINT - EPSG:4326
Row11	Ankara	Turkey	79.62	no	yes	39.93	32.86	9.86	POINT - EPSG:4326
Row12	Antalya	Turkey	79.62	no	yes	36.89	30.7	11.88	POINT - EPSG:4326
Row13	Arad	Romania	19.37	yes	yes	46.17	21.32	9.32	POINT - EPSG:4326
Row14	Athens	Greece	10.92	yes	yes	37.98	23.73	17.41	POINT - EPSG:4326
Row15	Augsburg	Germany	80.68	yes	yes	48.35	10.9	4.54	POINT - EPSG:4326
Row16	Bacau	Romania	19.37	yes	yes	46.58	26.92	7.51	POINT - EPSG:4326
Row17	Badajoz	Spain	46.06	yes	yes	38.88	-6.97	15.61	POINT - EPSG:4326
Row18	Baia Mare	Romania	19.37	yes	yes	47.66	23.58	8.87	POINT - EPSG:4326
Row19	Balti	Moldova	4.06	no	no	47.76	27.91	8.23	POINT - EPSG:4326
Row20	Barcelona	Spain	46.06	yes	yes	41.38	2.18	15.78	POINT - EPSG:4326
Row21	Bari	Italy	59.8	yes	yes	41.11	16.87	15.15	POINT - EPSG:4326
Row22	Basel	Switzerland	8.38	no	no	47.58	7.59	6.68	POINT - EPSG:4326
Row23	Batman	Turkey	79.62	no	yes	37.89	41.14	14.16	POINT - EPSG:4326
Row24	Belfast	United Kingdom	65.11	yes	yes	54.6	-5.96	8.48	POINT - EPSG:4326
Row25	Belgrade	Serbia	8.81	no	no	44.82	20.47	9.85	POINT - EPSG:4326
Row26	Bergamo	Italy	59.8	yes	yes	45.7	9.67	9.12	POINT - EPSG:4326
Row27	Bergen	Norway	5.27	no	yes	60.39	5.32	1.75	POINT - EPSG:4326

Figure 68: A Snapshot of the Resulting Geometry

c. Column Resorting: Columns were resorted into the preferred order for the map visualization by using the column resort node.

Row ID	EU	country	city	population	temperature	coastline	latitude	longitude	geometry
Row0	yes	Denmark	Aalborg	5.69	7.52	yes	57.03	9.92	POINT - EPSG:4326
Row1	yes	United Kingdom	Aberdeen	65.11	8.1	yes	57.17	-2.08	POINT - EPSG:4326
Row2	yes	Sweden	Abisko	9.85	0.2	yes	63.35	18.83	POINT - EPSG:4326
Row3	no	Turkey	Adana	79.62	18.67	yes	36.99	35.32	POINT - EPSG:4326
Row4	yes	Spain	Albacete	46.06	12.62	yes	39	-1.87	POINT - EPSG:4326
Row5	yes	Spain	Algeciras	46.06	17.38	yes	36.13	-5.47	POINT - EPSG:4326
Row6	yes	France	Amiens	64.67	10.17	yes	49.9	2.3	POINT - EPSG:4326
Row7	yes	Netherlands	Amsterdam	16.98	8.93	yes	52.35	4.92	POINT - EPSG:4326
Row8	yes	Italy	Ancona	59.8	13.52	yes	43.6	13.5	POINT - EPSG:4326
Row9	no	Andorra	Andorra	0.07	9.6	no	42.5	1.52	POINT - EPSG:4326
Row10	yes	France	Angers	64.67	10.98	yes	47.48	-0.53	POINT - EPSG:4326
Row11	no	Turkey	Ankara	79.62	9.86	yes	39.93	32.86	POINT - EPSG:4326
Row12	no	Turkey	Antalya	79.62	11.88	yes	36.89	30.7	POINT - EPSG:4326
Row13	yes	Romania	Arad	19.37	9.32	yes	46.17	21.32	POINT - EPSG:4326
Row14	yes	Greece	Athens	10.92	17.41	yes	37.98	23.73	POINT - EPSG:4326
Row15	yes	Germany	Augsburg	80.68	4.54	yes	48.35	10.9	POINT - EPSG:4326
Row16	yes	Romania	Bacau	19.37	7.51	yes	46.58	26.92	POINT - EPSG:4326
Row17	yes	Spain	Badajoz	46.06	15.61	yes	38.88	-6.97	POINT - EPSG:4326
Row18	yes	Romania	Baia Mare	19.37	8.87	yes	47.66	23.58	POINT - EPSG:4326
Row19	no	Moldova	Balti	4.06	8.23	no	47.76	27.91	POINT - EPSG:4326
Row20	yes	Spain	Barcelona	46.06	15.78	yes	41.38	2.18	POINT - EPSG:4326
Row21	yes	Italy	Bari	59.8	15.15	yes	41.11	16.87	POINT - EPSG:4326
Row22	no	Switzerland	Basel	8.38	6.68	no	47.58	7.59	POINT - EPSG:4326
Row23	no	Turkey	Batman	79.62	14.16	yes	37.89	41.14	POINT - EPSG:4326
Row24	yes	United Kingdom	Belfast	65.11	8.48	yes	54.6	-5.96	POINT - EPSG:4326
Row25	no	Serbia	Belgrade	8.81	9.85	no	44.82	20.47	POINT - EPSG:4326
Row26	yes	Italy	Bergamo	59.8	9.12	yes	45.7	9.67	POINT - EPSG:4326
Row27	no	Norway	Bergen	5.27	1.75	yes	60.39	5.32	POINT - EPSG:4326

Figure 69: A snapshot of the outcome of Colour Resorting.

d. Column Filter: In this section of the workflow, the dataset required for the visualization were filtered to the 'included' part of the setting while the columns that were not needed were left at the excluded part of the setting.

Table "default" - Rows: 213 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	EU	country	city	population	temperature	geometry
Row0	yes	Denmark	Aalborg	5.69	7.52	POINT - EPSG:4326
Row1	yes	United Kingd...	Aberdeen	65.11	8.1	POINT - EPSG:4326
Row2	yes	Sweden	Abisko	9.85	0.2	POINT - EPSG:4326
Row3	no	Turkey	Adana	79.62	18.67	POINT - EPSG:4326
Row4	yes	Spain	Albacete	46.06	12.62	POINT - EPSG:4326
Row5	yes	Spain	Algeciras	46.06	17.38	POINT - EPSG:4326
Row6	yes	France	Amiens	64.67	10.17	POINT - EPSG:4326
Row7	yes	Netherlands	Amsterdam	16.98	8.93	POINT - EPSG:4326
Row8	yes	Italy	Ancona	59.8	13.52	POINT - EPSG:4326
Row9	no	Andorra	Andorra	0.07	9.6	POINT - EPSG:4326
Row10	yes	France	Angers	64.67	10.98	POINT - EPSG:4326
Row11	no	Turkey	Ankara	79.62	9.86	POINT - EPSG:4326
Row12	no	Turkey	Antalya	79.62	11.88	POINT - EPSG:4326
Row13	yes	Romania	Arad	19.37	9.32	POINT - EPSG:4326
Row14	yes	Greece	Athens	10.92	17.41	POINT - EPSG:4326
Row15	yes	Germany	Augsburg	80.68	4.54	POINT - EPSG:4326
Row16	yes	Romania	Bacau	19.37	7.51	POINT - EPSG:4326
Row17	yes	Spain	Badajoz	46.06	15.61	POINT - EPSG:4326
Row18	yes	Romania	Baia Mare	19.37	8.87	POINT - EPSG:4326
Row19	no	Moldova	Balti	4.06	8.23	POINT - EPSG:4326
Row20	yes	Spain	Barcelona	46.06	15.78	POINT - EPSG:4326
Row21	yes	Italy	Bari	59.8	15.15	POINT - EPSG:4326
Row22	no	Switzerland	Basel	8.38	6.68	POINT - EPSG:4326
Row23	no	Turkey	Batman	79.62	14.16	POINT - EPSG:4326
Row24	yes	United Kingd...	Belfast	65.11	8.48	POINT - EPSG:4326
Row25	no	Serbia	Belgrade	8.81	9.85	POINT - EPSG:4326
Row26	yes	Italy	Bergamo	59.8	9.12	POINT - EPSG:4326
Row27	no	Norway	Bergen	5.27	1.75	POINT - EPSG:4326

Figure 70: Snapshot of the Target Variables

5.3 Map Visualizations Generated Using the Workflow

Four visualizations were carried out using the workflow, 2 on heat map (Population and Temperature and 2 on Geospatial view (Population and Temperature).

a. Population (Heat Map)

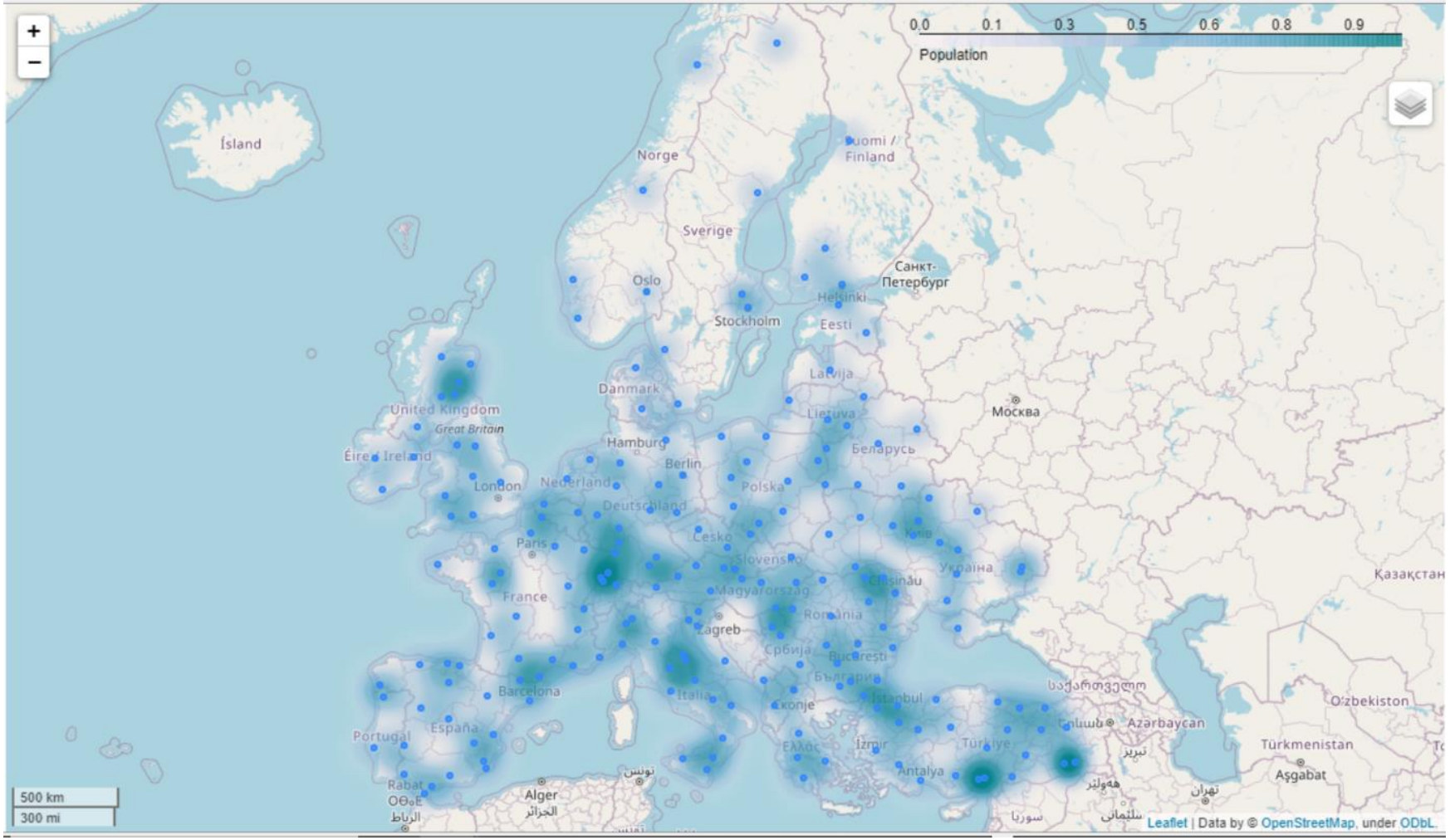


Figure 71: Temperature of European and Non-European Cities.

b. Temperature (Heat Map)

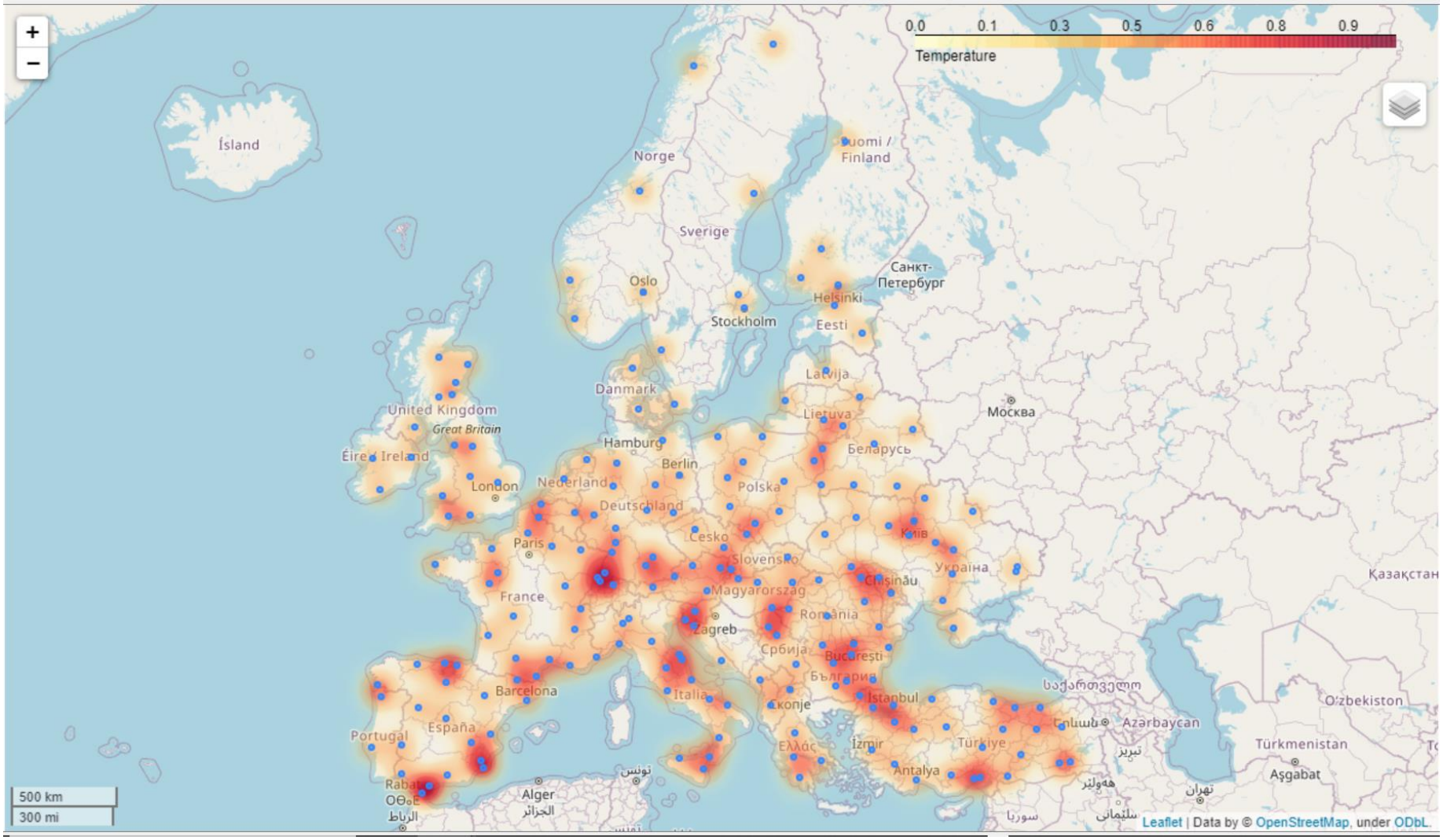


Figure 72: Temperature of European and Non-European Cities

Figures 71 and 72 were heat maps representing the visualizations of temperature and population of some European and Non-European cities according to the dataset. These visuals were generated using the spatial heatmap node. This type of visualization is called Choropleth map. It is good in that it is easy to read and understand however, it has some shortcomings. It's makes visual salience to depend on region size i.e. not a true representation of the attribute values population and temperature and the colour palette choice has a huge influence on the resulting visuals.

Based on the shortcomings of the Choropleth Map, an alternative map was considered which was Symbol Map with the following output.

c. Population (Symbol Map)

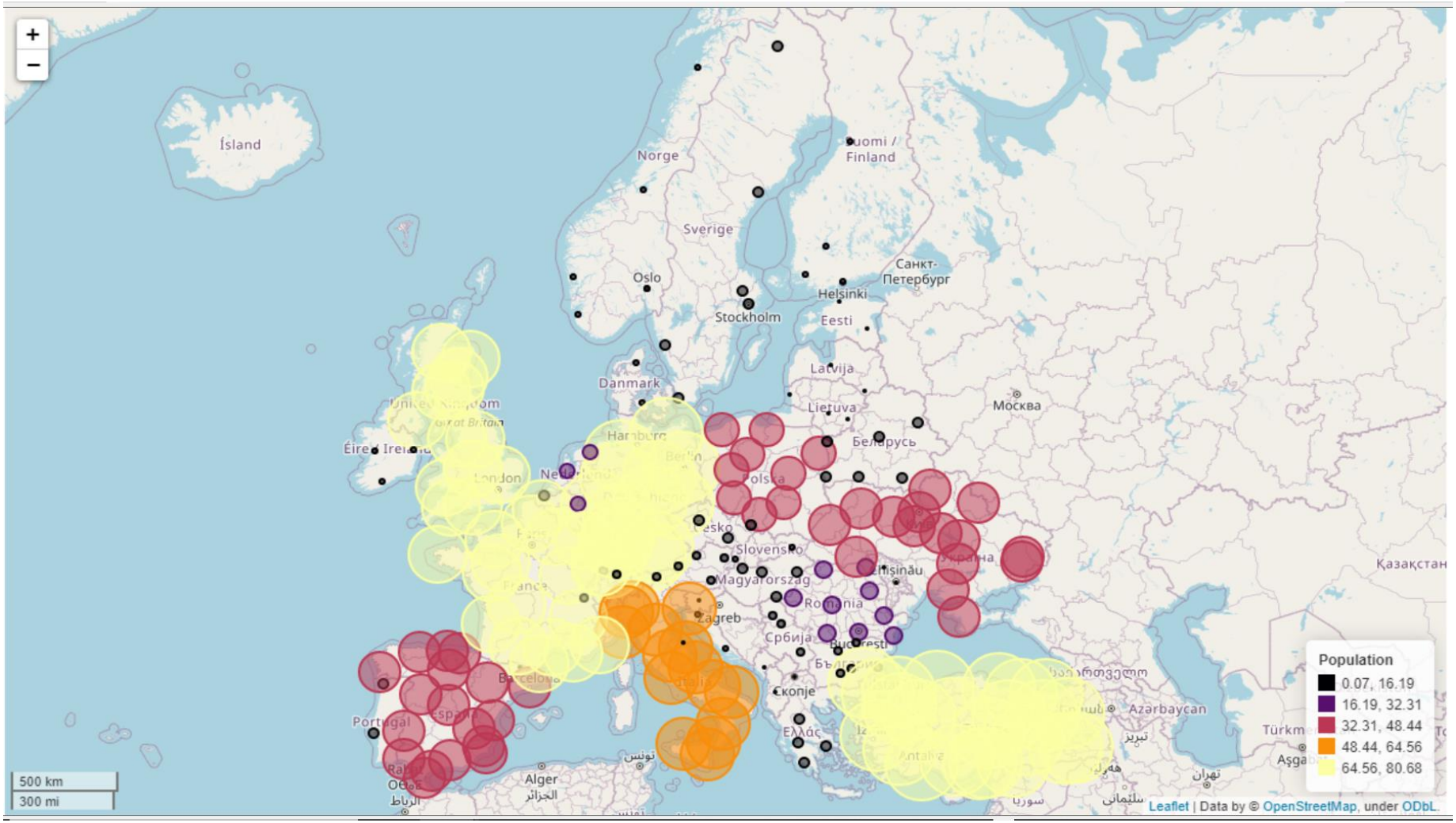


Figure 73: Population of Some European and Non-European Cities

d. Temperature (Symbol Map)

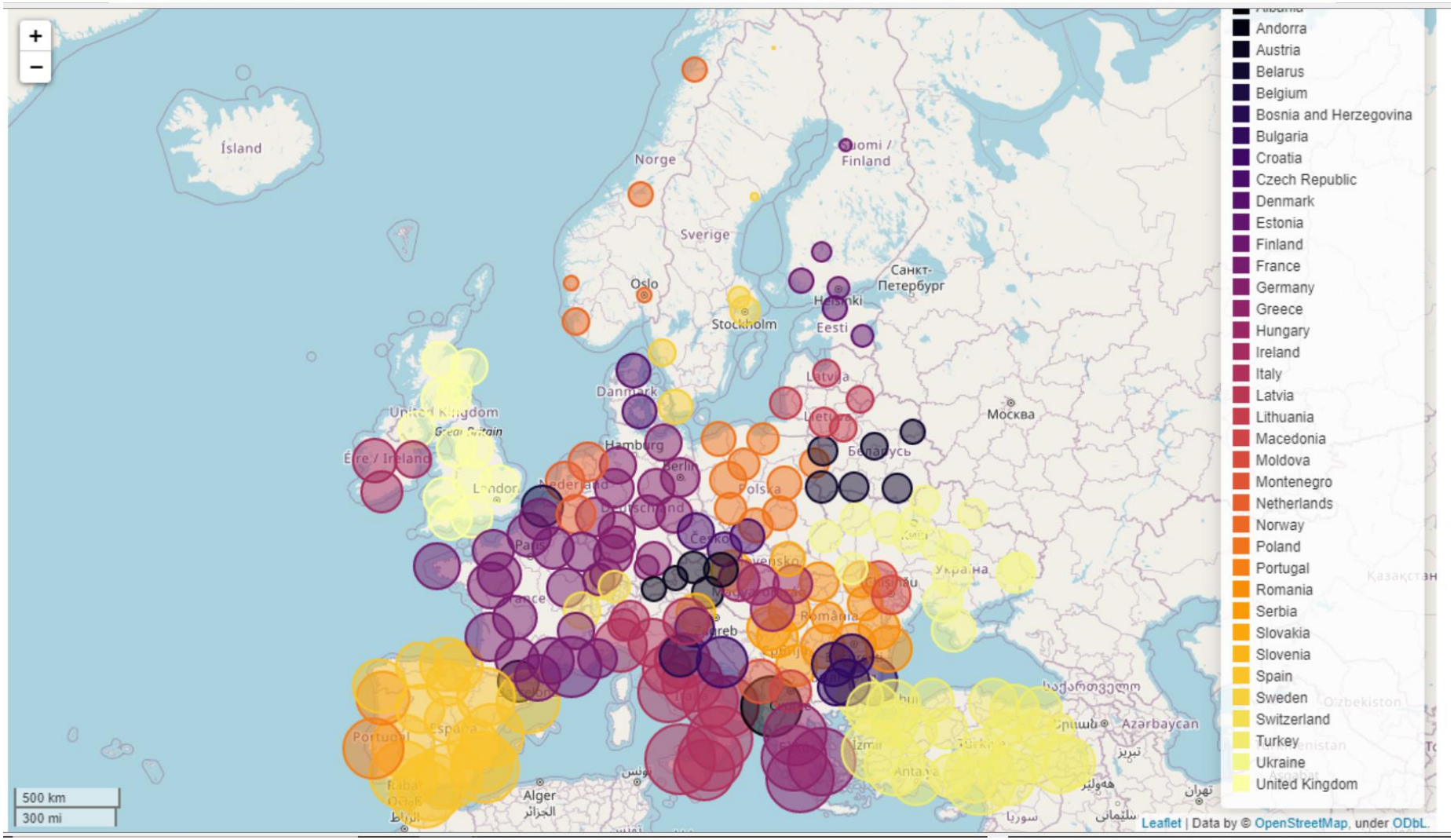


Figure 74: Temperature of Some European and Non-European Cities

Figures 73 and 74 were symbol maps representing the visualizations of population and temperature of some European and Non-European cities according to the dataset. These visuals were generated using the 'Geospatial View Node'. This type of visualization is called Symbol map. It is good a better option in comparison to choropleth maps for the following reasons:

i. It allows use of size and shape and color channels
ii. It keep original spatial geometry in the background and often a good alternative to choropleth maps
iii. It is easy to read and understand thereby eliminating the problems with region size vs data salience
iv. Marks: symbol size is a true representation of attribute value and they are uniform
v. glyphs: symbol size can be uniform

Symbol maps are not perfect options of map visual because of the reasons:

i. There is possible congestion or overlap i.e. symbols overlap each other thereby causing a possible obstruction of the other regional boundaries.

6.0 Conclusion

There exists several data analytics visualization methods and tools with each them having their strengths and weaknesses. However, Knime analytics is more of a robust data science tool which has several advantages and functionalities over others. It stands out by the following offerings:

1. It's a one single, open source data analytics tool for Building of Analytical Workflows Using an Intuitive User-Interface.
2. Scale Execution with Demands: It can scale workflow performance through in-memory streaming and multi-threaded data processing.
3. It exercises the power of in-database processing or distributed computing on Apache Spark to further increase computation performance.
4. It can blend data from any source using different nodes in the knime repository.
5. It can blend different data types (strings, integers, images, text, networks, sound, molecules, and more).
6. It can connect to all major databases and data warehouses such as SQL Server, Postgres, MySQL, Snowflake, Redshift, BigQuery, and more.
7. Blend large data volumes: It has the functionality to import and export HDFS data and perform SQL analytics within Hive and Impala, or create and run Apache Spark applications within KNIME.

Appendix

Table of Figures

References

Amy G., 2015. Harvard Business Review. A Refresher on Regression Analysis. (Available Online at): https://hbr.org/2015/11/a-refresher-on-regression-analysis, Accessed on: 19th March 2023.

CS 102: Working with Data Tools and Techniques - Spring 2020. (Available Online): https://web.stanford.edu/class/cs102/datasets.htm, Accessed on 20th March 2023.

J. Thomas and K. Cook, editors. Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE Press, 2005. (Available Online): https://www.osti.gov/biblio/912515, Accessed on 15th March 2023.

Kaggle.com 2023. (Available Online): https://www.kaggle.com/datasets/HRAnalyticRepository/employee-attrition-data. Accessed on: March 15th 2023.

Keim, D., Kohlhammer, J., Ellis, G. and Mansmann, F., 2010. Mastering the information age: solving problems with visual analytics. (Available on-line at): https://kops.uni-konstanz.de/bitstream/handle/123456789/12737/VisMaster-Book_127373.pdf?sequence=2, Accessed on 15th March 2023

KNIME Network Visualization. (Available Online at): https://www.knime.com/book/network-visualization. Accessed on: 20th March 2023

KNIME Spatial Data Visualizationhttps://hub.knime.com/knime/extensions/org.knime.features.ext.osm/latest. Accessed on: 16th March 2023

Knime, 06 Aggregations. (Available Online at): https://hub.knime.com/knime/spaces/Education/latest/Self-Paced%20Courses/Archive/L1-DS%20KNIME%20Analytics%20Platform%20for%20Data%20Scientists%20-%20Basics%20(deprecated)/Exercises/06%20Aggregations~2eu0Ckw92jpUtgLu , Accessed on 15th March 2023.

Knime, 05 Data Visualization. (Available Online at): https://hub.knime.com/knime/spaces/Education/latest/Self-Paced%20Courses/Archive/L1-DS%20KNIME%20Analytics%20Platform%20for%20Data%20Scientists%20-%20Basics%20(deprecated)/Exercises/05%20Data%20Visualization~pUo_k7WCaXUxvw3D, Accessed on: 14th March 2023.

Knime, 08 Regression Model. (Available Online at): https://hub.knime.com/knime/spaces/Education/latest/Self-Paced%20Courses/Archive/L1-DS%20KNIME%20Analytics%20Platform%20for%20Data%20Scientists%20-%20Basics%20(deprecated)/Exercises/08%20Regression%20Model~VKIbW-eaS5L43Qhi, Accessed on: 15th March 2023.

Knime, 03 Row and Column Filtering (Available Online): https://hub.knime.com/knime/spaces/Education/latest/Self-Paced%20Courses/Archive/L1-DS%20KNIME%20Analytics%20Platform%20for%20Data%20Scientists%20-%20Basics%20(deprecated)/Exercises/03%20Row%20and%20Column%20Filtering~gxvz2UDhSoiZ3NHD , Accessed on: March 14th 2023.

Knime, Univariate Visual Exploration with Data Explorer node. (Available Online): https://hub.knime.com/knime/spaces/Examples/latest/03_Visualization/02_JavaScript/11_Univariate_Visual_Exploration_with_Data_Explorer~Z1bIrOvg409Ok8GP , Accessed on: 14th March 2023.

Leishi Zhang, 2023 Visual Data Analytics, MSc in Data Intelligence Data Visual Analytics

Milan J., Misleading Data Visualization Examples to Stay Away From. (Available Online): https://wpdatatables.com/misleading-data-visualization-examples/ , Accessed on: 13th March 2023.

Mohiuddin O., 2022. Bad Data Visualization Examples: Fix It or Risk It. Available Online: https://ninjatables.com/bad-data-visualization-examples/, Accessed on: 10th March, 2023.

Mohiuddin O., 2023. Data Visualization Basics, Skills & Techniques. Available Online: https://ninjatables.com/data-visualization/ , Accessed on: 10th March 2023.

P. C. Wong and J. Thomas. Visual analytics. IEEE Computer Graphics and Applications, 24(5):20–21, 2004.

TosinLitics, Introduction to 30 Days of Knime. (Available Online): https://www.youtube.com/watch?v=R4TDey5px4U&list=PLrVumvjxxTkAs1o4sXxnqXpFeAQTgTjmV , Accessed on: 3rd March 2023.