2021

Project 2

CASE STUDIES IN CLASSIFICATION, REGRESSION, AND TRIAGE

BUS 212A Analyzing Big Data II

Jianying Zhu

May 10, 2021

Table of Contents

1 Introduction

1.1 Introduction of Report

This project focus on 3 main types of the supervised modeling: regression, classification, and triage, which are targeting 3 different datasets that called *Retail Data Analytics, Diabetes Data Set, and World Happiness Report.* The entire modeling process includes Data Preparation, Data Reduction and Descriptive Analysis, Modelings and Conclusion and Suggestions.

Here is some relevant background information for datasets:

The original Retail Data Analytics datasets has 3 files that provided the historical sales data of 45 stores located in different regions that each store includes a number of departments. Throughout the year, it also runs several promotional markdown events that precede important holidays, the 4 largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. In these files, holidays evaluation for a week is weighted five times higher than non-holiday weeks.

Diabetes is a chronic and metabolic disorder disease and even deadly, which hampers a common man's life at the extreme end. It also leads to many other diseases, such as blindness, blood pressure, heart disease, and kidney disease, and liver damage. It also is a major public health challenge worldwide. According to a study by the World Health Organization (WHO), the number of diabetes will have raised to 552 million by 2030, denote that one in ten grownups will have diabetes if no serious measure is taken.

The World Happiness Report is a landmark survey of the status of global happiness. It contains articles, and rankings of national happiness based on respondent ratings of their own lives on that 0 to 10 scales, which the report also correlates with various life factors. The happiness scores and rankings use data from the Gallup World Poll. Our report analysis starts with the third publication dataset in 2015 and the following 4 years, which ranks 155 countries by their happiness levels, released at the United Nations. These publications continually gain global recognition as governments, organizations, and civil society increasingly uses happiness indicators to inform their policy-making decisions. This report will assess the national and regional variations in well-being.

In the Data Preparation, this report goes through aggregating variables, combing tables, extracting important variables into separate columns, and dealing with missing data and outliers. After cleaning, this report contains 20 variables with 421570 rows for the Retails Dataset, 9 variables with 768 rows for the Diabetes Dataset, 18 variables with 727 rows for the World Happiness Report Dataset.

In the Descriptive Analysis, this report uses tables and correlation matrix to identify the relationships between target variable and candidate predictor variables. This report also applies the principal components technique to reduce dimension of each dataset.

In the Modeling Performance, this report develops several supervised learning algorithms for each dataset, comparing and selecting the best performance models. This report also chooses some unseen data to test the best models chosen for each dataset.

In the last part as Conclusion and Suggestions, this report interprets the result of best model and reflects on the entire modeling process. Finally, this report provides some insights and advice for the hypothesis as shown below.

1.2 Research Questions and Hypothesis

What are you asking of the dataset to solve? What domain questions are you trying to answer?

Overall, I want to deal with the answers of 3 questions below for each dataset:

- Figure out important factors in all variables and predict the department-wide sales for the following year
- This report will try to predict the presence of diabetes based on some relevant covariates and find out some important factors that caused diabetes.
- Regarding the World Happiness, this report will compare the situation between both developed and developing countries and will discuss the factors affect people happy among all the countries.

2 Data Preparation

2.1 Data Source and Description

In this report, all the analysis is based on different sub-datasets. They all from 3 initial datasets, *Retail Data Analytics, Diabetes Data Set, and World Happiness Report* that are collected from Kaggle, published by Manjeet Singh[1], National Institute of Diabetes and Digestive and Kidney Diseases[2], and Sustainable Development Solutions Network[3].

First of all, the original Retail Data Analytics datasets contain 3 datasets: “features dataset.csv”(12 columns, 8190 rows), “sales dataset.csv”(5 columns, 421570 rows), and “stores dataset.csv”(3 columns, 45 rows).

- Feature dataset contains additional data related to the store, department, and regional activity for the given dates during that week.

[1] https://www.kaggle.com/manjeetsingh/retaildataset
[2] https://www.kaggle.com/mathchi/diabetes-data-set
[3] https://www.kaggle.com/unsdsn/world-happiness

- Sales dataset covers 3 years from 2010-02-05 to 2012-11-01, department numbers, and conditional variables representing if a holiday week or not.
- Store dataset indicates the type and size of the store.

Second, the original Diabetes Datasets only contain 1 dataset, “diabetes.csv”(9 columns, 768 rows), taken from a larger database. The objects of this dataset are the Pima Indian heritage females whose age are equal to or greater than 21. The reason that the participants were selected is that the incidence of diabetes is higher. This dataset consists of 9 attributes with 8 attribute values and 1 class variable including two outcomes, namely whether the patient is tested positive (indicated by output 1) or tested negative (indicated by 0).

Lastly, the original World Happiness Report contain 5 dataset, “2015.csv”(12 columns, 158 rows), “2016.csv”(13 columns, 157 rows), “2017.csv”(12 columns, 155 rows), “2018.csv” (9 columns, 156 rows), and “2019.csv”(9 columns, 156 rows). This dataset consists of 15 attributes with 13 attribute values and 2 category variables including country and region.

2.2 Data Cleaning and Data Quality

Regarding the data process, it is a technique of machine learning that converts raw data into a logical or comprehensible format. An important method is data cleaning. This section involves certain activities, like integrating the data, checking for duplicate values, missing values, and type mismatches. All these inconsistencies will be eliminated from datasets. It is very important to clean the datasets before training them on a classifier in order to better learn the hidden pattern in the datasets. This process guarantees that the three primary datasets maintain data correctness and integrity with high quality. It is ready for further analysis.

Table 1: Attributes of Retails Data Set

	Variable Name	Variable Type	Variable Description
1	Store	Integer	Number of stores which observation in recorded 1-45
2	Dept	Integer	Number of departments ranging from 1-99
3	Date	Datatime	Time of the week where this observation was taken
4	Weekly Sales	Integer	Sales for the given department in the given store($)
5	Is Holiday	Boolean	Whether the week is a special holiday week(True, False)
6	Temperature	Numeric	Average temperature of the region during that week(°F)
7	Fuel Price	Numeric	Costs of fuel in the region($)
8	MarkDown1-5	Numeric	Types of markdowns and quantities available during that week. Only available after Nov 2011, but not available for all stores all the time
9	CPI	Numeric	Consumer Price Index
10	Unemployment	Numeric	Unemployment rate
11	Type	Object	3 types(A, B, C)
12	Size	Integer	Calculated by the numbers of products available in the particular store ranging from 34,000 to 210,000(sq. ft.)

For the retail dataset, two files of initial datasets, sales, and features, are merged by 3 unique variables: store, date, and IsHoliday. This data frame then merges with the store table by the unique store variable. Most of predicted variables do not have missing values, but Markdowns 1-5 have 270889,

310322, 284479, 286603, and 270138 values respectively. This dataset fills in NaN values with a mean of each column to remove those missing values.

Table 2: Attributes of Diabetes Data Set

	Variable Name	Variable Type	Variable Description
1	Pregnancies	Integer	Number of times pregnant
2	Glucose	Integer	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
3	Blood Pressure	Integer	Diastolic Blood Pressure (mm Hg)
4	Skin Thickness	Integer	Triceps skin fold thickness(mm)
5	Insulin	Integer	2-hour serum insulin((µU/ml)
6	BMI	Numeric	Body Mass Index(kg/m2)
7	Diabetes Pedigree Function	Numeric	History of diabetes in relatives or generic
8	Age	Integer	Age(years)
9	Outcome	Integer	Occurrence of Diabetes (0 or 1)

For the diabetes dataset, missing values does not exist when checking at first. It appeared after writing NaN values of 0 for Glucose, Blood Pressure, Skin Thickness, Insulin, BMI. There are abnormal or zero values for the variables: glucose, blood pressure, skin thickness, Insulin, and BMI, represents 5, 35, 227, 374 and 11 values respectively. To remove all of them, this dataset fills in NaN values with a median according to the target variable outcome.

Table 3: Attributes of World Happiness Report Data Set

	Variable Name	Variable Type	Variable Description
1	Country	Category	Name of the Countries ranked by values of happiness
2	Region	Category	Regions of the listed countries
3	Year	Integer	2015-2019
4	Happiness Rank	Integer	Rank of the country based on the Happiness Score
5	Happiness Score	Numeric	A metric measured by rate the happiness on a scale of 0 to 10 where 10 is the happiest
6	Standard Error	Numeric	A measure of margins of error in distributions
7	Economy (GDP per Capita)	Numeric	The extent to which GDP contributes to the calculation of the Happiness Score
8	Family	Numeric	The extent to which Family contributes to the calculation of the Happiness Score
9	Health (Life Expectancy)	Numeric	The extent to which Life expectancy contributed to the calculation of the Happiness Score
10	Freedom	Numeric	The extent to which Freedom contributed to the calculation of the Happiness Score
11	Trust (Government Corruption)	Numeric	The extent to which Perception of Corruption contributes to Happiness Score
12	Generosity	Numeric	The extent to which Generosity contributed to the calculation of the Happiness Score
13	Dystopia Residual	Numeric	The extent to which Dystopia Residual contributed to the calculation of the Happiness Score
14	Upper Confidence Interval(Whisker High)	Numeric	Lower Confidence Interval of the Happiness Score
15	Lower Confidence Interval(Whisker Low)	Numeric	Upper Confidence Interval of the Happiness Score

The 5 world happiness report datasets are checked separately at first. 3 datasets of 2017, 2018 and 2019 renamed columns to: 'Happiness Rank', 'Happiness Score', 'Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)', 'Freedom', and 'Trust (Government Corruption)' which matched the column names in

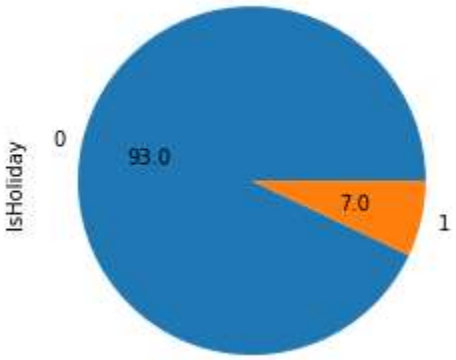
2015 and 2016. The column of Country combined the names of country and region in the datasets of 2017, 2018 and 2019, in order to separate them, these 3 datasets mapped regions that are not available, and added the column of Regions to these countries. The next step is to check the values of 'Trust (Government Corruption)' in these 3 datasets because only one missing value appeared in 2018 dataset, so it set an average value between 2017 and 2019 for this dataset. No missing values are shown for each 5 datasets, then they are combined. The combined datasets also checked the missing value. To fill the null or missing value of the columns of Country and Region, this dataset renamed some countries, removed unrecognized countries, mapped missing regions, and changed the type of object to category. The datasets do not consider the columns of Lower Confidence Interval, Upper Confidence Interval, high Whisker and low Whisker because they don't provide useful information in this report.

2.3 Data Distribution

This section mainly presents data visualization for each of 3 major datasets, including pie chart, histogram, and boxplots. They are fully displayed the changing trends of different predicted variables, and the relationship between them and target variables. Detailed information will be discussed below.

2.3.1 *Retails Data Set*

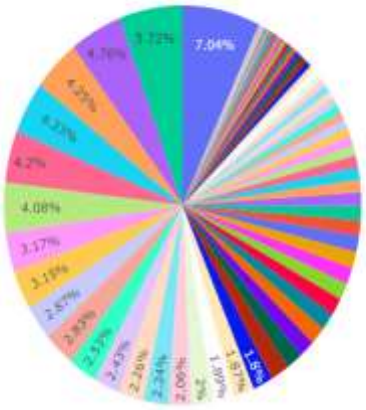
Figure 1-1: Pie charts: Holiday



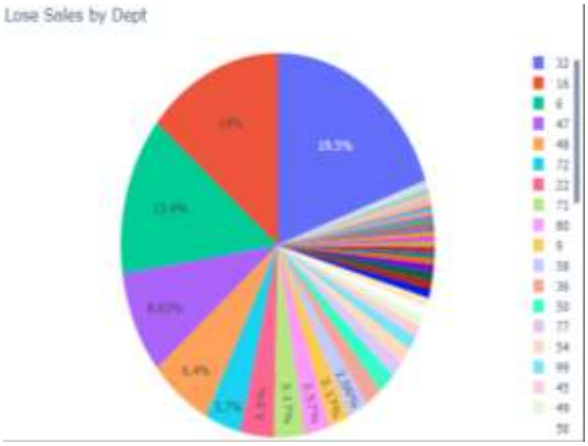
The dataset has 392,060 nonholidays and 295,099 holidays. This sample is highly unbalanced with a 93% of non-holiday classified as 0 versus 7% of holiday records classified as 1 in Diabetes.

Figure 1-2: Pie charts: Department and Stores

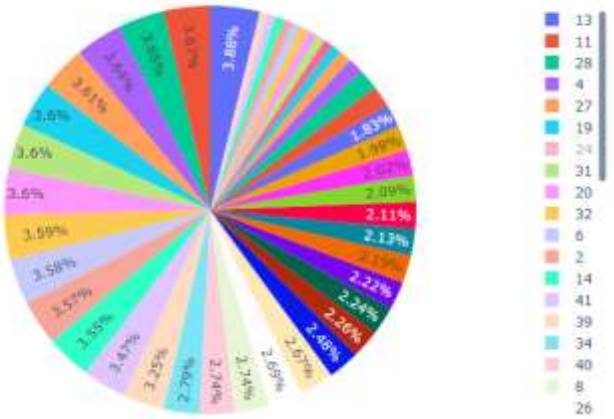
Profit Sales by Dept



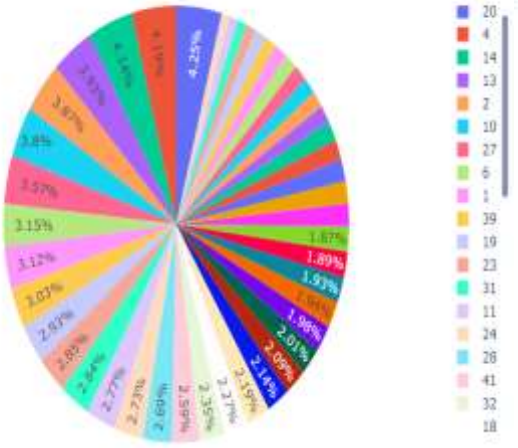
Lose Sales by Dept



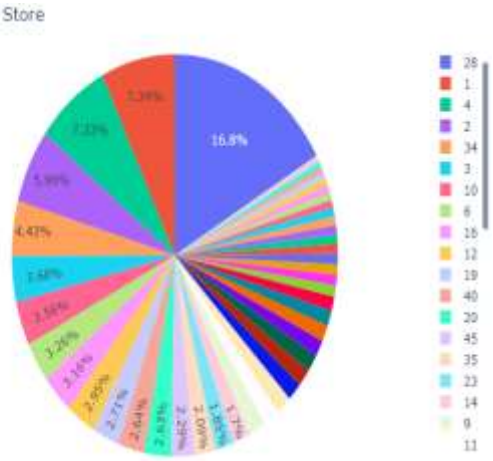
Store by Size



Profit Sales by Store



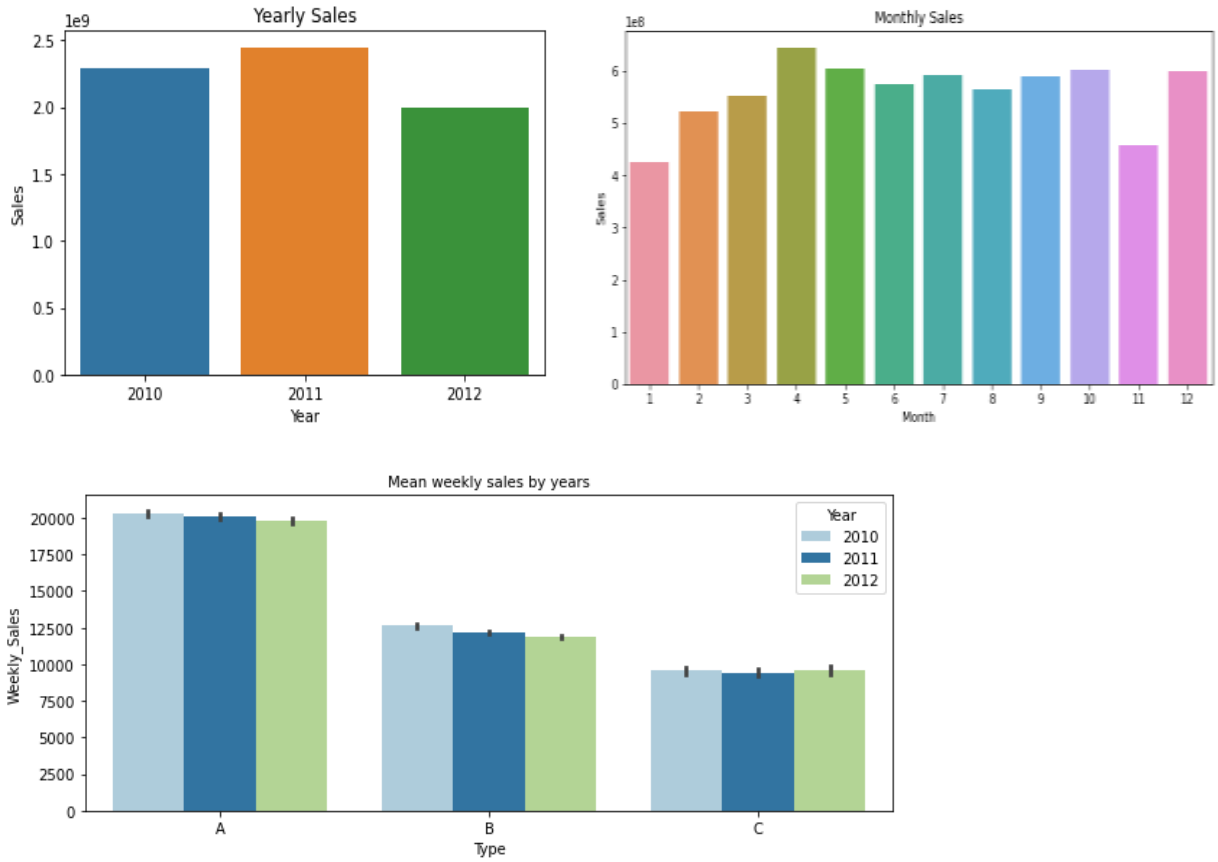
Lose Sales by Store



These pie charts shows the distribution by department and by stores. Amongst all 99 departments, it shows the dept92 generate the most profit sales around 7.04%, but the dept 43 has the least profit sales. The dept32 generate the highest loss around 19.5%. These 45 stores are different by sizes, store 13 takes up to 3.88% but store 5 takes up the least the proportion. Store 20 generates about 4.25%

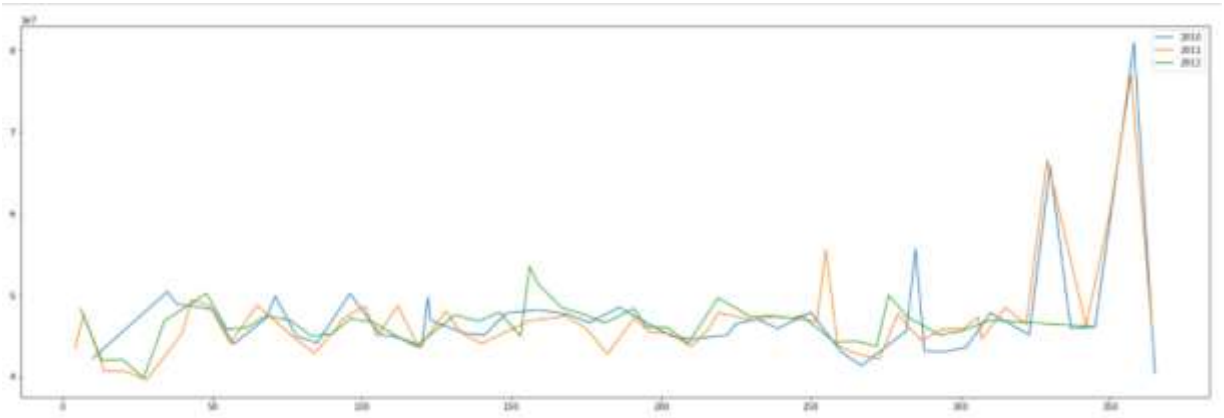
profit sales, but store 5 has the least one. Store 28 has the highest losses about 16.8%, and store 30 has the lowest sales.

Figure 2: Bar Graph

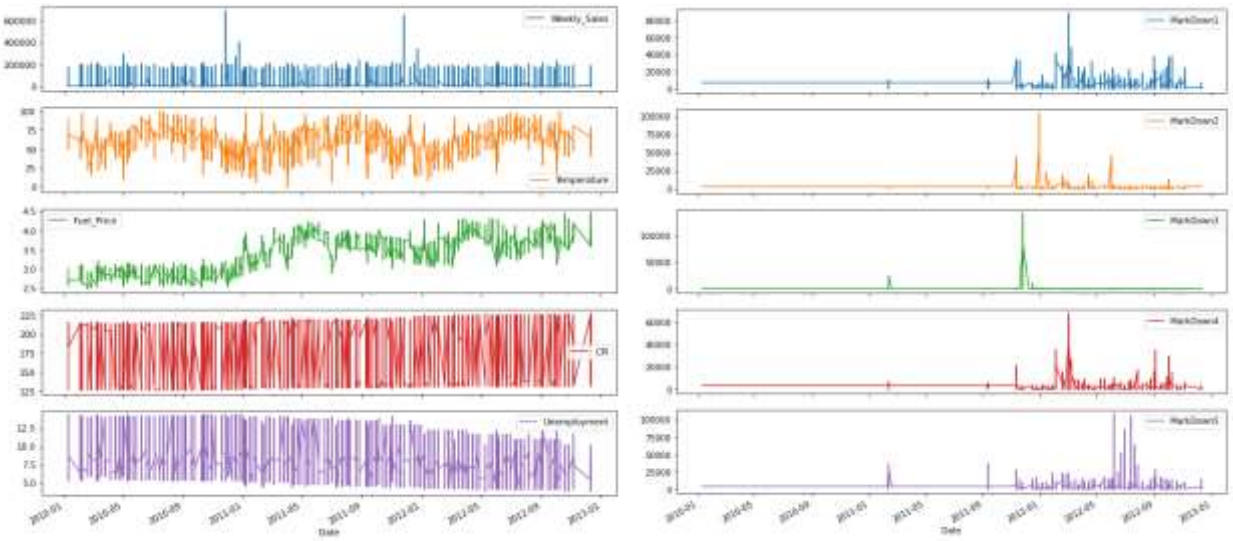


These bar plots show sales by year and by month. Store Type A and B has shown slight decrease year on basis in sales, but store C has shown a slight increase in sales.

Figure 3: Line Chart

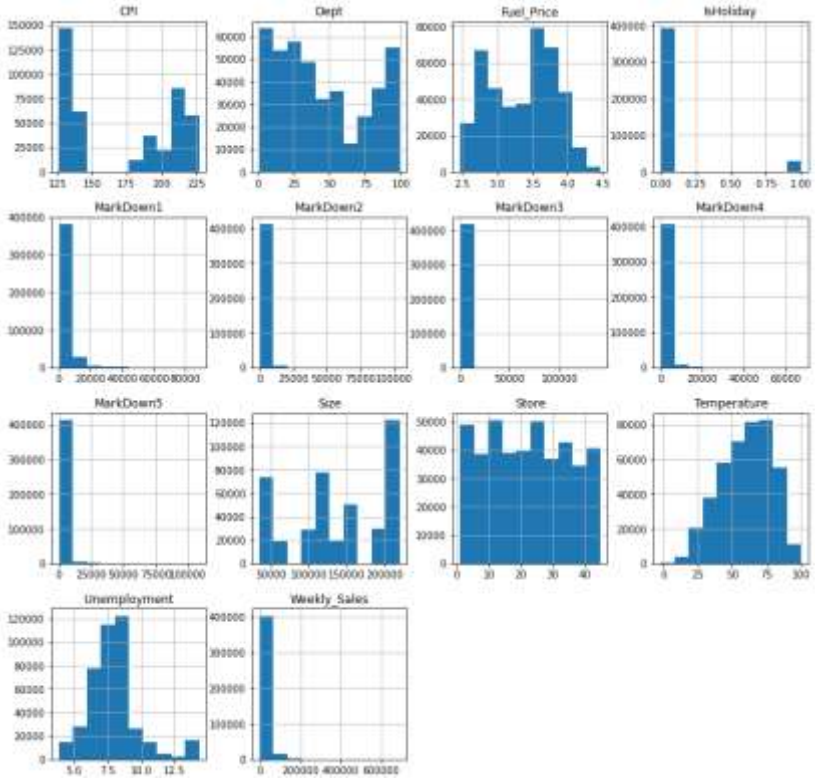


This line chart presents a time series of weekly sales from 2010 to 2012.



This line graph shows the upward and downward trends of above 10 factors from 2010 to 2012. It displays large numbers of sales during the entire year of 2012.

Figure 4: Histograms



In the histogram above, Unemployment is right-skewed distributions, but Temperature is left-skewed distributions.

Figure 5: Boxplots

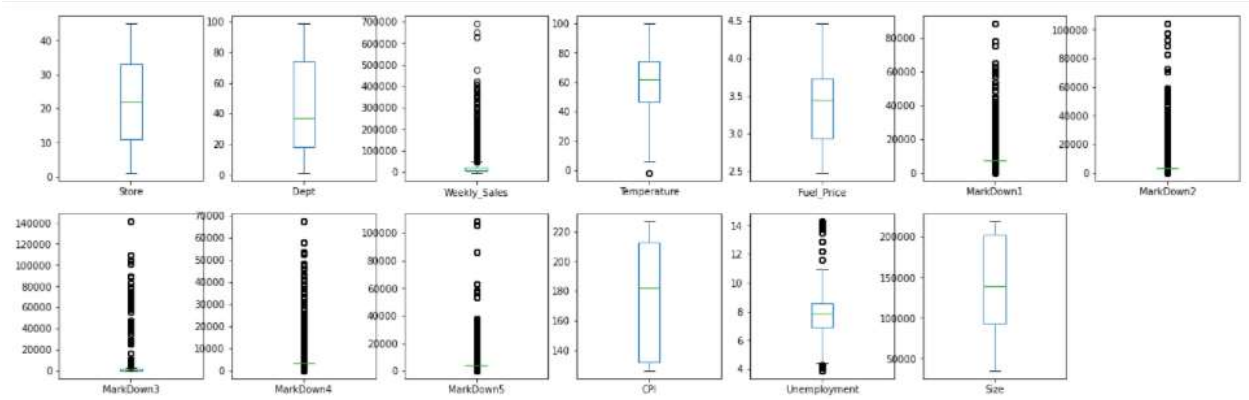
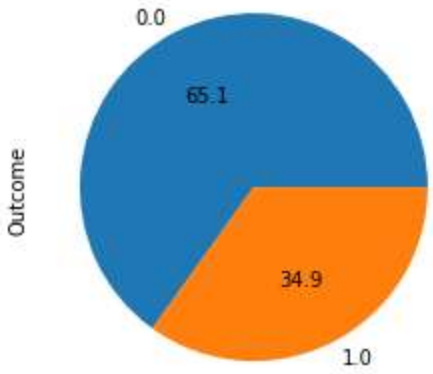


Figure 5 shows the boxplots of the explanatory variables. Department and Unemployment show little variation.

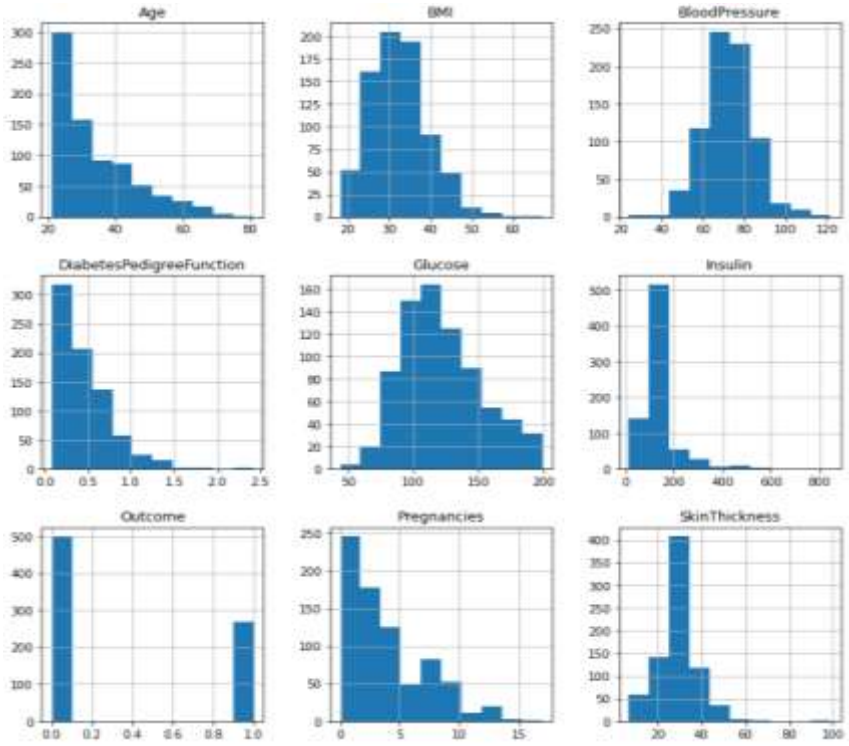
2.3.2 Diabetes Data Set

Figure 1: Pie chart



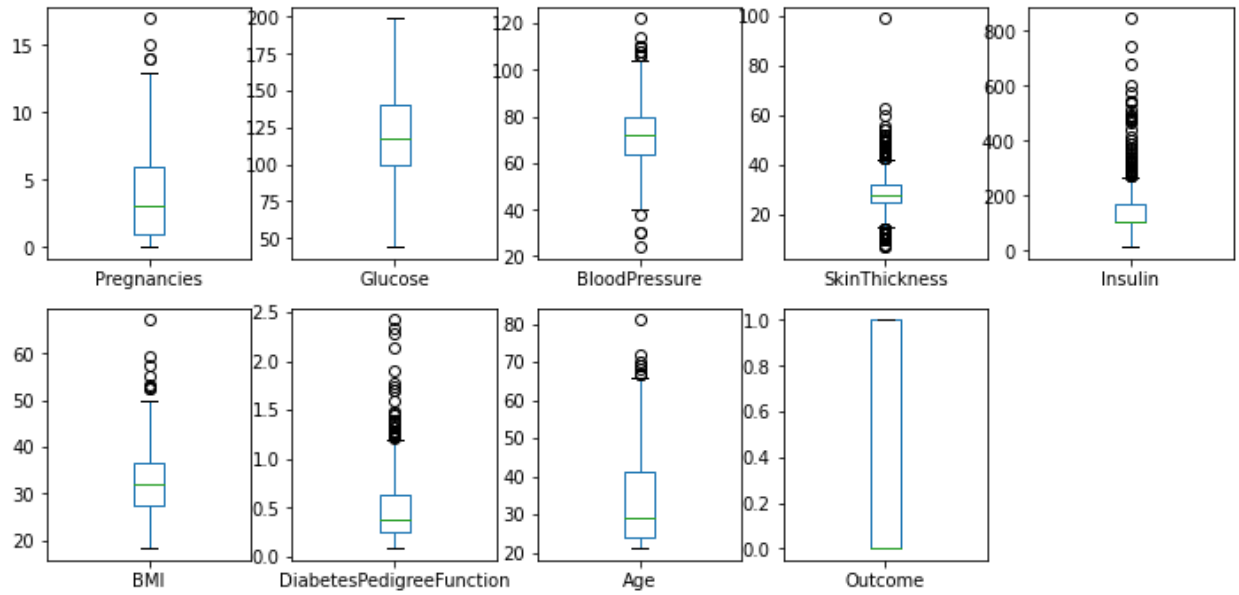
The dataset has 268 women that were diagnosed with Diabetes and 500 women that did not have Diabetes. This sample is highly unbalanced with a 65.1% of non-outcome classified as 0 versus 34.9% of outcome records classified as 1 in Diabetes.

Figure 2: Histograms



From figure 2 of the histogram plots, it is evident that the variables of Pregnant and Age are highly skewed. The plots of Glucose and blood pressure are similar to the normal distribution. For continuous variable, this report will get more clarity on the distribution by analyzing it with the dependent variables.

Figure 3: Boxplots



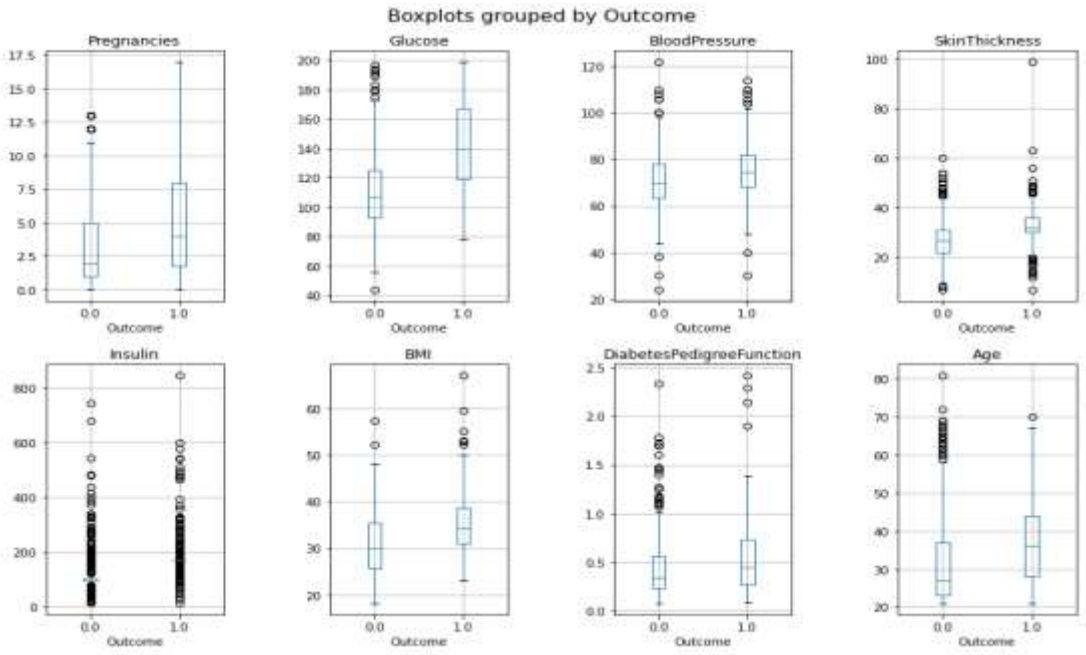
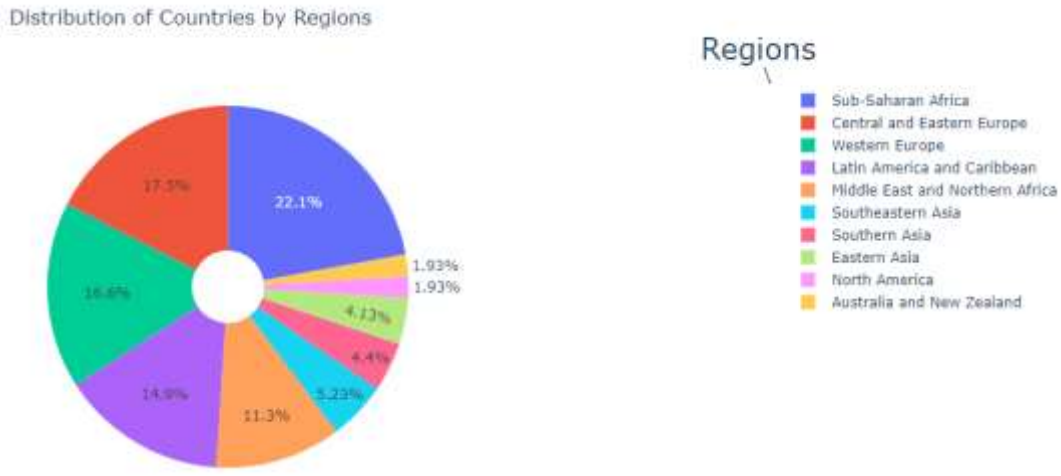


Figure 3 shows the boxplots of the explanatory variables. It gives an idea about the features of the datasets, including 8 predicted variables and target variable - outcome. These plots show that several columns have outliers, such as blood pressure, skin thickness, and BMI, with the column Insulin being the most critical. From the analysis, this report could infer that median glucose content is higher for patients who have diabetes. Blood pressure and skin thickness show little variation with the diabetes.

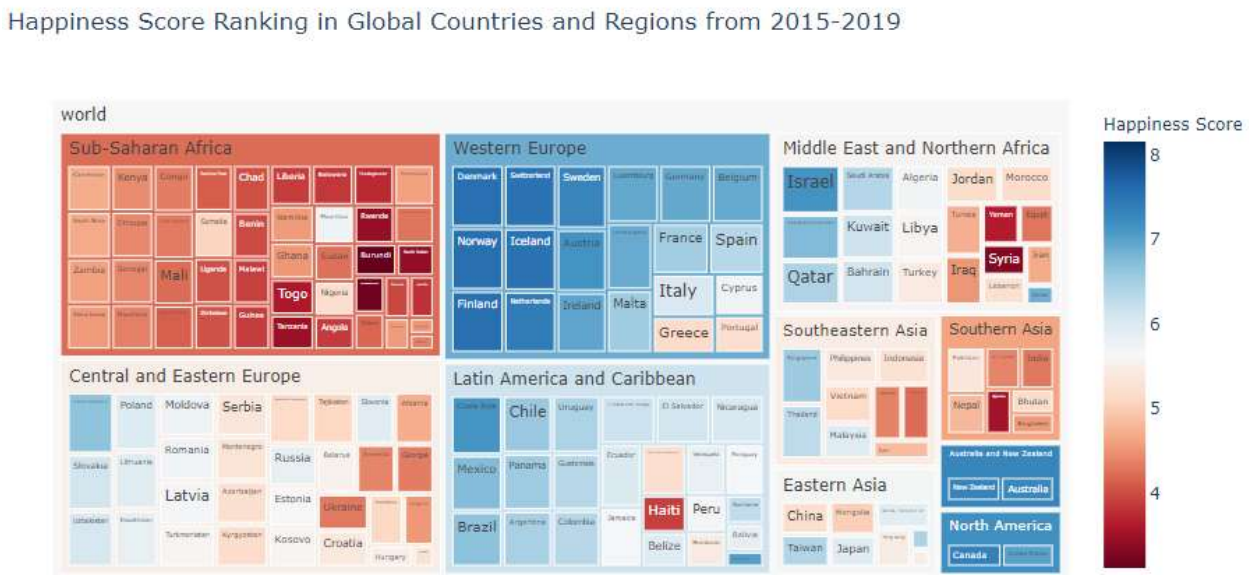
2.3.3 World Happiness Report Data Set

Figure 1: Pie chart



This pie chart presented 10 distribution among all the countries and regions, and it calculated the distribution of countries are in each region. Sub-Saharan Africa and Central and Eastern European are the most distribution, but the Pacific Ocean countries like Australia and New Zealand are the least distribution.

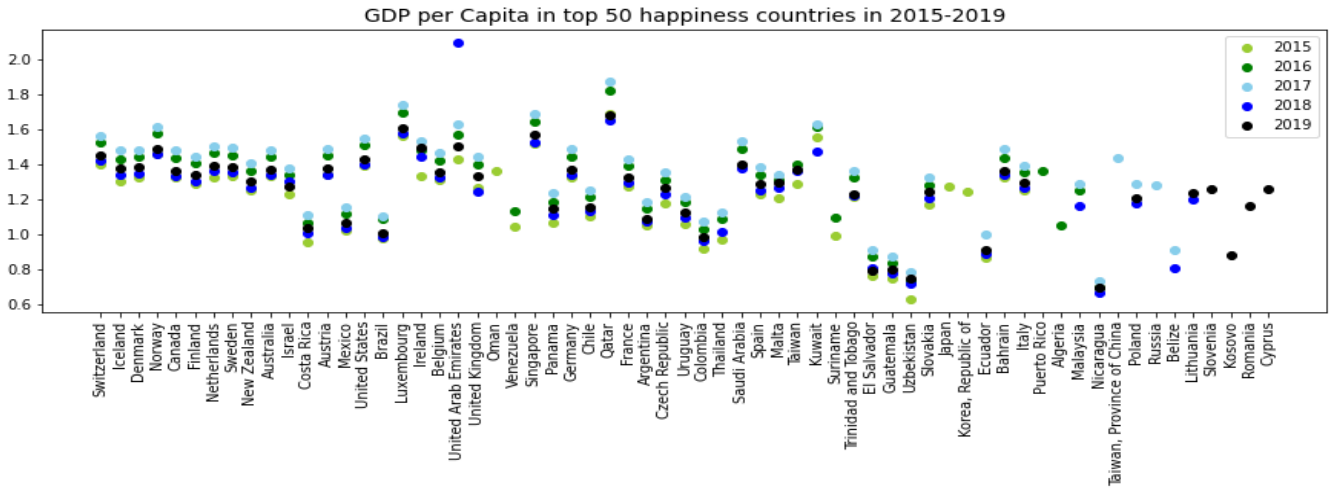
Figure 2: Tree map Chart



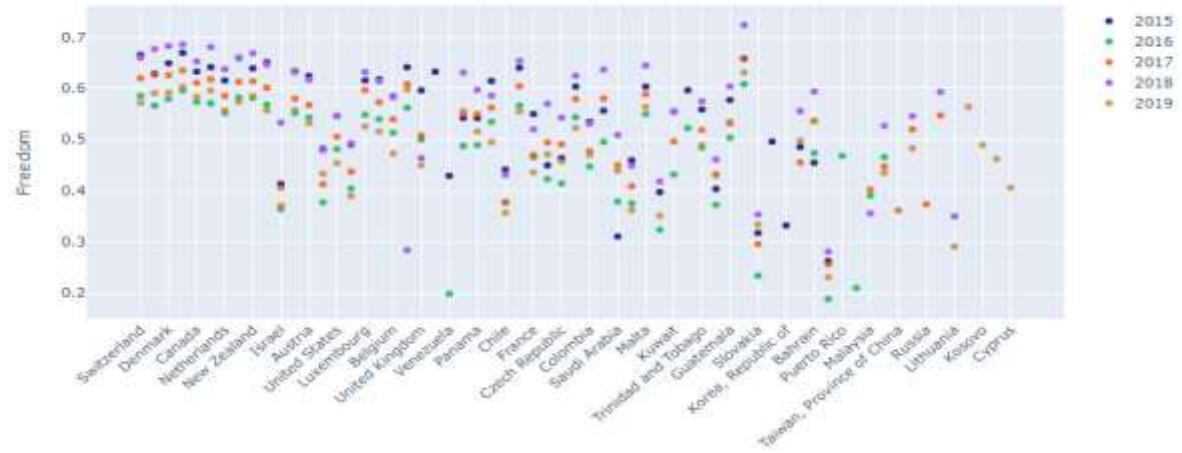
Note that each tree map chart of 2015 – 2019 are shown in the relevant coding file.

Red countries are seeming the most unhappy countries on above map. These unhappy countries are also from same regions, African Countries are always in the bottom levels. The Situations are very bad for Sub-Saharan Region as it is the unhappiest region in the world. Northern Europe, North America, Canada, Australia are on the top. They are the Happiest Regions so far.

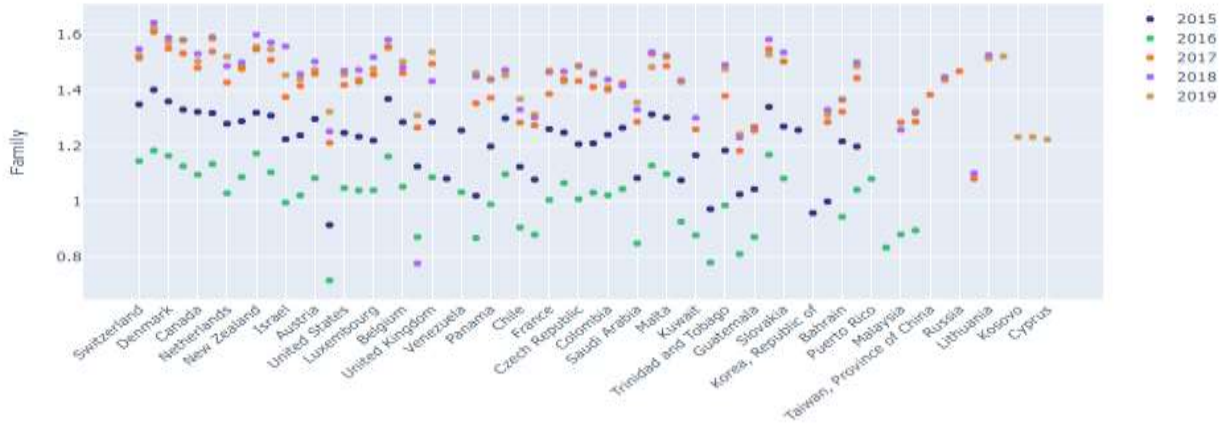
Figure 3: A group of Scatter plots



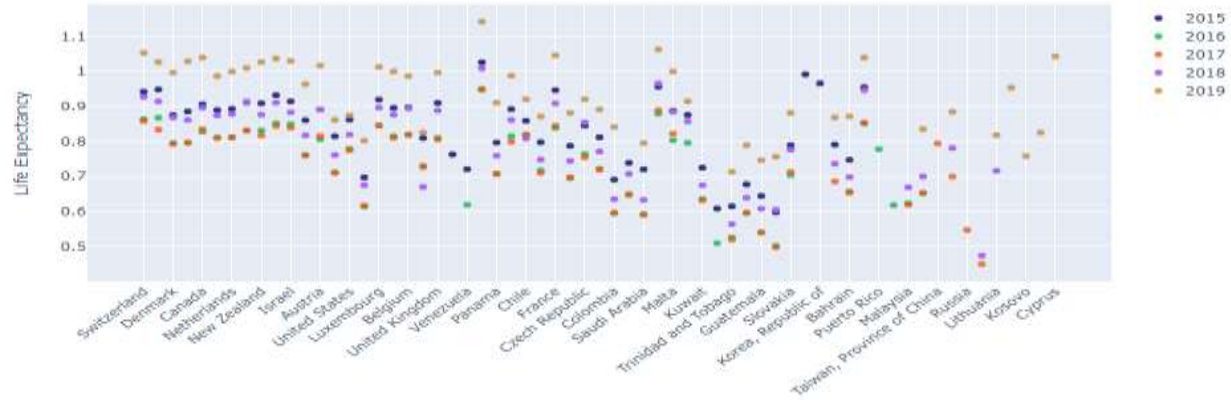
Freedom vs Happiness Rank of Top 50 Countries in 2015-2019 Years

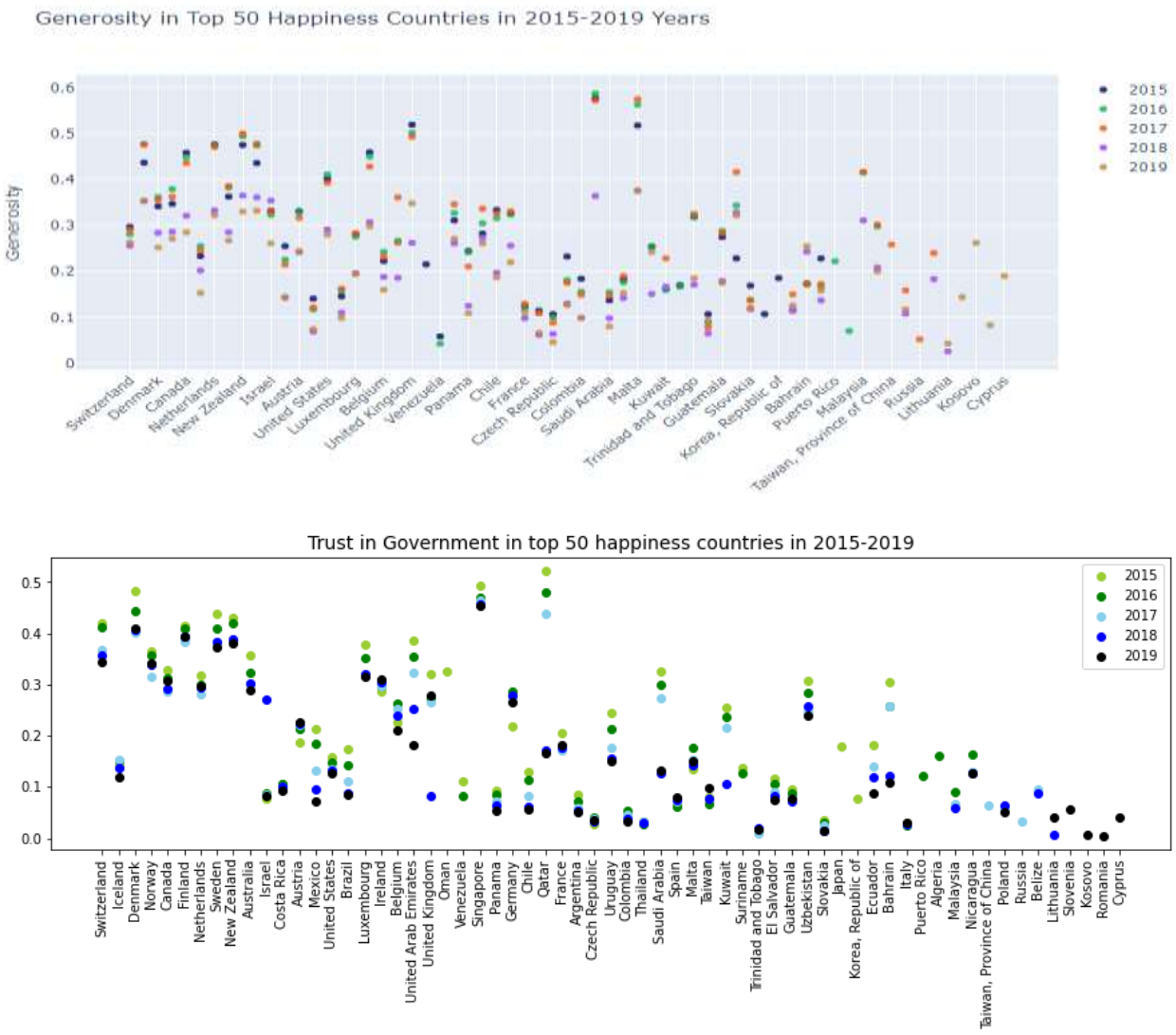


Family support in Top 50 Happiness Countries in 2015-2019 Years



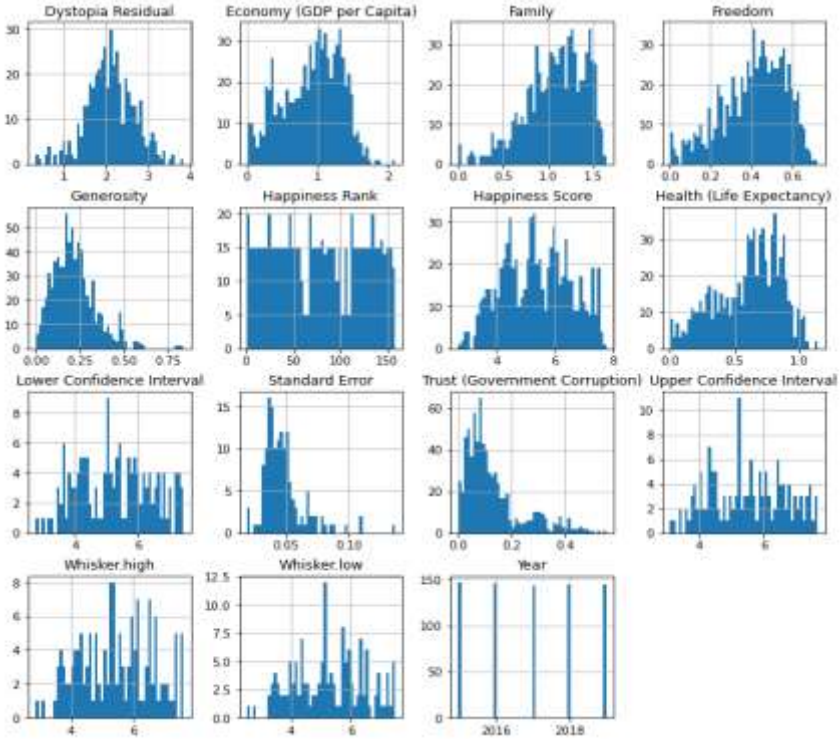
Life Expectancy in Top 50 Happiness Countries in 2015-2019 Years





These scatter plots presented 50 happiness countries affecting by 6 factors: Economy(GDP), Generosity, Family Satisfaction, Life Expectation, Freedom and Trust in Government, from 2015 to 2019. Northern Europe, North America, Canada, Australia stands at position number 50 amongst the Happiness Rankings for the World. They are able to keep their people happy. The correlations are quite good with almost all the important factors being highly correlated with Happiness. Some factors like Economy(GDP), Freedom, and Family Satisfaction, are dynamic as they are mostly important factors for happiness.

Figure 4: Histograms



In the histogram above, the variables of Generosity, Trust in Government and standard error are highly skewed. Economy(GDP) has a little right-skewed distributions. But the variables of Family Satisfaction, Freedom and Health(Life Expectations) are highly left-skewed distributions. Besides, the distribution of Dystopia Residual is roughly symmetric, and the values distributed between approximately 0 and 4.

Figure 5: Boxplots

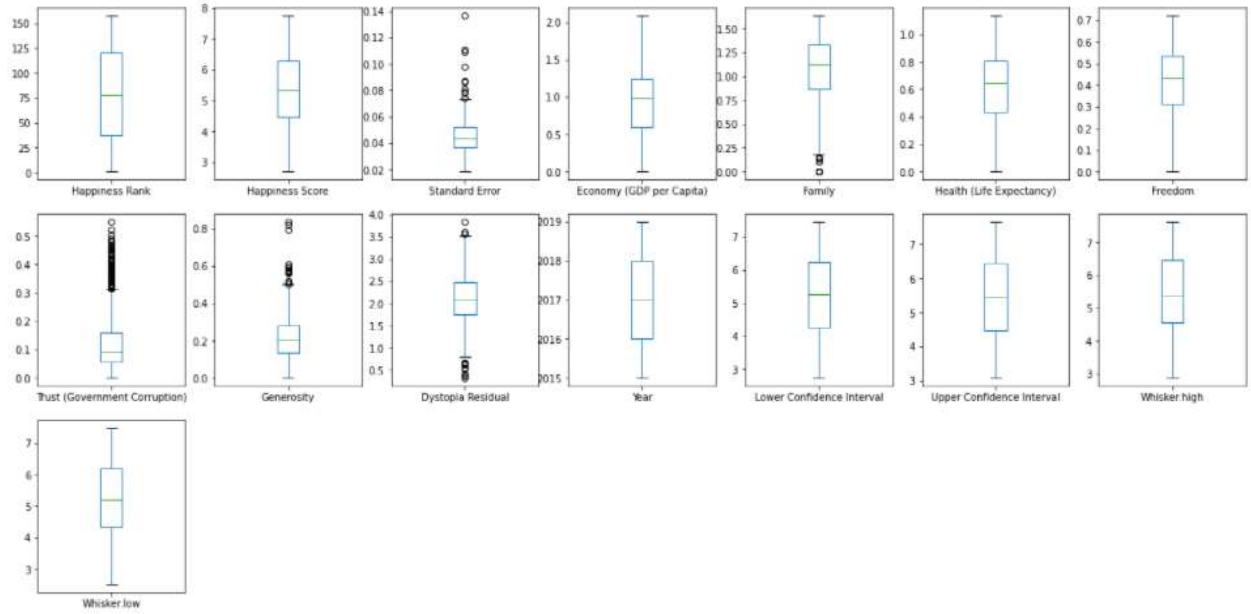


Figure 5 also shows the boxplots of the explanatory variables. These plots show that several columns have outliers, including Family Satisfaction, Trust in Government, Generosity, Dystopia Residual and standard error. These variables also have lowest variation.

3 Descriptive Analysis and Data Reduction

3.1 Descriptive Statistics

The detailed descriptive statistics of important variables for each dataset are listed in the table below, including mean, standard deviation, median, minimum, maximum.

Table 4: Descriptive Statistics of Retails Data Set

	count	mean	std	min	5%	25%	50%	75%	90%	95%	99%	max
Store	421570.0	22.200546	12.785297	1.000	3.000000	11.000000	22.000000	33.000000	40.000000	43.000000	45.000000	45.000000
Dept	421570.0	44.260317	30.492054	1.000	4.000000	18.000000	37.000000	74.000000	92.000000	95.000000	98.000000	99.000000
Weekly_Sales	421570.0	15981.258123	22711.183519	-4988.940	58.974500	2079.650000	7612.030000	20205.852500	42845.673000	61201.951000	105479.586000	693099.360000
Size	421570.0	136727.915739	60980.583328	34875.000	39690.000000	93638.000000	140167.000000	202505.000000	204184.000000	206302.000000	219622.000000	219622.000000
Temperature	421570.0	60.090059	18.447931	-2.060	27.310000	46.680000	62.090000	74.280000	83.580000	87.270000	92.810000	100.140000
Fuel_Price	421570.0	3.361027	0.458515	2.472	2.653000	2.933000	3.452000	3.738000	3.917000	4.029000	4.202000	4.468000
MarkDown1	421570.0	7246.420196	4956.920816	0.270	709.320000	7246.420196	7246.420196	7246.420196	8624.560000	12497.710000	28177.290000	88646.760000
MarkDown2	421570.0	3334.628621	4867.493911	-265.760	22.800000	3334.628621	3334.628621	3334.628621	3334.628621	3789.560000	21813.150000	104519.540000
MarkDown3	421570.0	1439.421384	5487.601593	-29.100	2.700000	115.380000	1439.421384	1439.421384	1439.421384	1439.421384	2083.280000	141630.610000
MarkDown4	421570.0	3383.168256	3560.353127	0.220	273.530000	3383.168256	3383.168256	3383.168256	3383.168256	5163.630000	15822.460000	67474.850000
MarkDown5	421570.0	4628.975079	3573.795304	135.160	1288.810000	4628.975079	4628.975079	4628.975079	5284.630000	7456.150000	15590.530000	108519.280000
CPI	421570.0	171.201947	39.159276	126.064	126.496258	132.022667	182.318780	212.416993	219.444244	221.941558	225.473509	227.232807
Unemployment	421570.0	7.960289	1.863296	3.879	5.326000	6.891000	7.866000	8.572000	9.816000	12.187000	14.180000	14.313000

Table 5: Descriptive Statistics of Diabetes Data Set

	count	mean	std	min	5%	25%	50%	75%	90%	95%	99%	max
Pregnancies	768.0	3.845052	3.369578	0.000	0.00000	1.00000	3.0000	6.00000	9.0000	10.00000	13.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	79.00000	99.00000	117.0000	140.25000	167.0000	181.00000	196.00000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	38.70000	62.00000	72.0000	80.00000	88.0000	90.00000	106.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	0.00000	23.0000	32.00000	40.0000	44.00000	51.33000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	0.00000	30.5000	127.25000	210.0000	293.00000	519.90000	846.00
BMI	768.0	31.992578	7.884160	0.000	21.80000	27.30000	32.0000	36.60000	41.5000	44.39500	50.75900	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.14035	0.24375	0.3725	0.62625	0.8786	1.13285	1.69833	2.42
Age	768.0	33.240885	11.760232	21.000	21.00000	24.00000	29.0000	41.00000	51.0000	58.00000	67.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.00000	0.0000	1.00000	1.0000	1.00000	1.00000	1.00

Table 6: Descriptive Statistics of World Happiness Report Data Set

	count	mean	std	min	5%	25%	50%	75%	90%	95%	99%	max
Happiness Rank	727.0	78.298487	46.494856	1.000000	8.000000	37.000000	78.000000	120.500000	142.000000	149.700000	155.000000	158.000000
Happiness Score	727.0	5.388019	1.161780	2.693000	3.577100	4.451500	5.332000	6.296500	6.993400	7.321600	7.549580	7.769000
Standard Error	147.0	0.047591	0.017199	0.018480	0.030909	0.036780	0.043610	0.051825	0.068092	0.078768	0.109884	0.136930
Economy (GDP per Capita)	727.0	0.914437	0.415043	0.000000	0.205272	0.583865	0.981240	1.247430	1.405282	1.489195	1.674900	2.096000
Family	727.0	1.079134	0.333476	0.000000	0.440000	0.870456	1.123236	1.331000	1.479840	1.525700	1.583740	1.644000
Health (Life Expectancy)	727.0	0.610337	0.250348	0.000000	0.158317	0.431392	0.647380	0.808579	0.893400	0.954838	1.038220	1.141000
Freedom	727.0	0.413060	0.154905	0.000000	0.121606	0.312000	0.434000	0.534660	0.596593	0.632000	0.669540	0.724000
Trust (Government Corruption)	727.0	0.128085	0.107366	0.000000	0.016405	0.056000	0.093000	0.159775	0.299594	0.376486	0.462408	0.551910
Generosity	727.0	0.220474	0.122455	0.000000	0.055000	0.134000	0.204435	0.280853	0.378044	0.470277	0.574053	0.838075
Dystopia Residual	437.0	2.096558	0.571674	0.328580	1.110570	1.749222	2.097538	2.465700	2.823716	3.054800	3.398611	3.637720
Year	727.0	2016.993122	1.417602	2015.000000	2015.000000	2016.000000	2017.000000	2018.000000	2019.000000	2019.000000	2019.000000	2019.000000
Lower Confidence Interval	146.0	5.290021	1.183780	2.732000	3.540000	4.260750	5.266000	6.223500	6.902500	7.258250	7.424850	7.480000
Upper Confidence Interval	146.0	5.488870	1.172010	3.078000	3.769500	4.454750	5.451000	6.462000	7.116500	7.391750	7.591100	7.669000
Whisker.high	144.0	5.461618	1.152940	2.864884	3.668031	4.550768	5.380905	6.446130	7.031961	7.376062	7.588977	7.622030
Whisker.low	144.0	5.263938	1.180601	2.521116	3.433605	4.326389	5.193687	6.197101	6.906440	7.244043	7.446773	7.479956

Note that 2015 – 2019 Descriptive Statistics tables are shown in the relevant coding file.

3.2 Correlation Analysis

Correlation analysis is the most common method of predictive regression. It assesses the degree of strength, direction of association, and a linear summary of relationship existing between two variables, or observational units (Berg, 2004). In an effort to expose the descriptive analysis, correlational patterns resulting from the dataset, this section shows confusion matrixes based on the tables of statistical analyses above, indicating that the role and relationships of independent and dependent variables, which is essential in the analysis of data.

Figure 6: Correlation Matrix of Retails Data Set

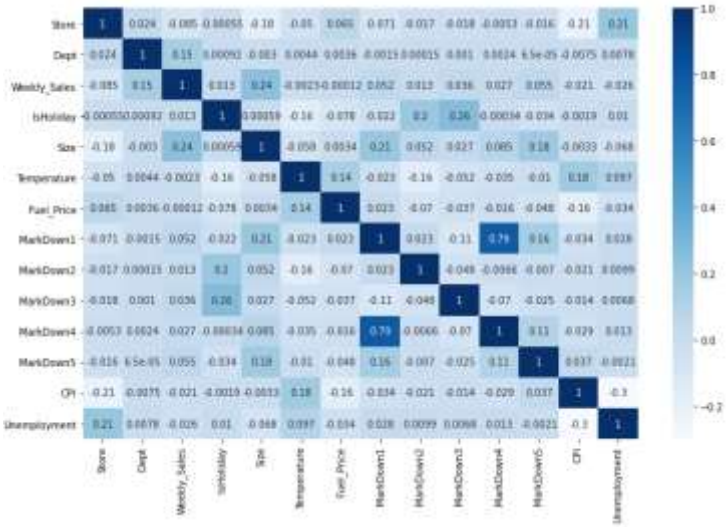
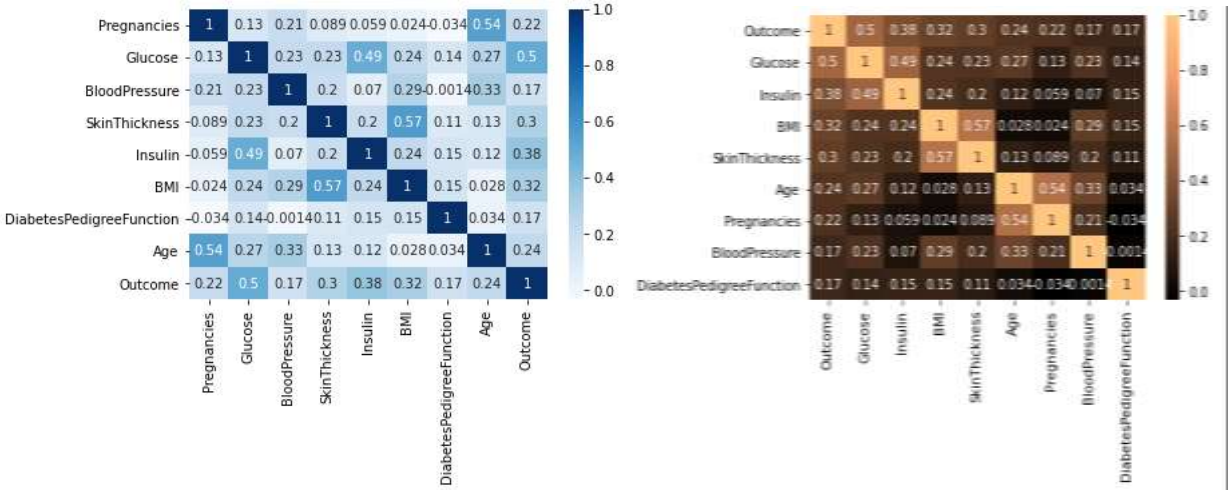


Figure 6 displays that there is a high correlation between the promotion markdown 1 and markdown 4. There is somewhat positive relationship between size, department, holiday and weekly sales, but these is negative relationship between temperature, CPI, fuel price, unemployment and weekly sales. Note that there is a little or even negative relations between the predicted variables.

Figure 7: Correlation Matrix of Diabetes Data Set

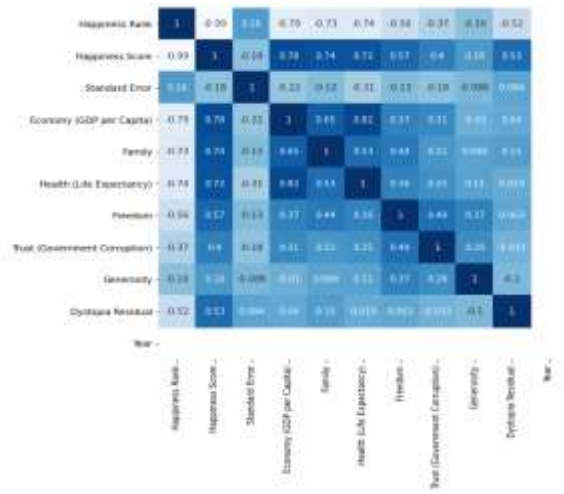
Left: for all variables *Right: for target variable Outcome*

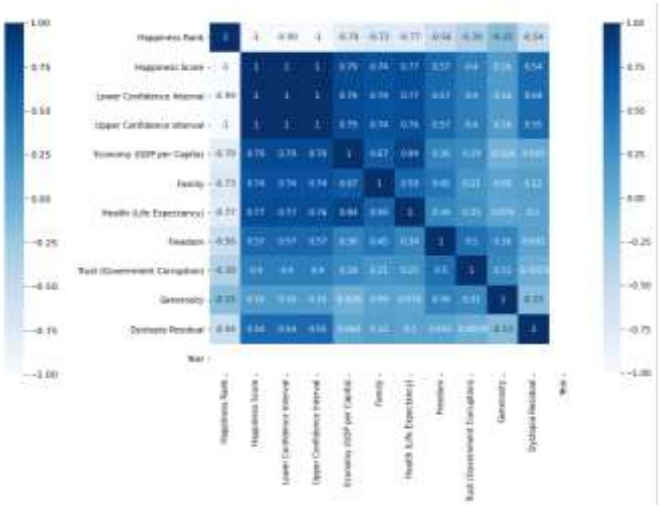


In figure 7, there is a little or no correlation exists between the predicted variables. Hence the model is not likely to suffer from multicollinearity. Insulin and Glucose, BMI, and Skin Thickness have a moderate to linear correlation. But there is a strong relationship between the predicted variables and the target. The highest correlation score of r = 0.5 suggests the strongest, positive relationship between the variables "glucose" and "outcome", whereas two correlations of r = 0.17 suggest weak associations between two predicted variables "blood pressure" and "Diabetes Pedigree Function" with "outcome".

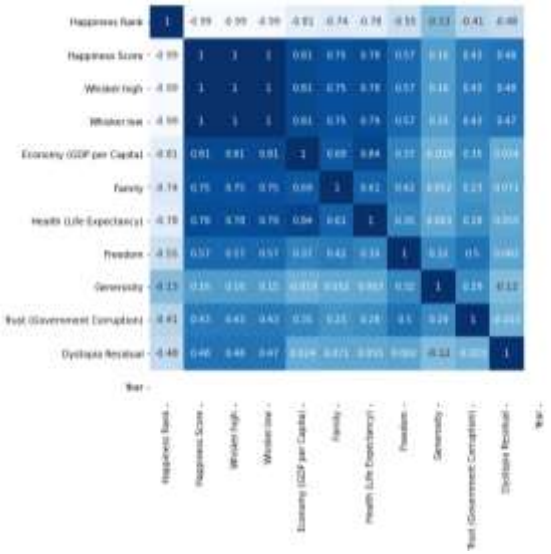
Figure 8: A Group of Correlation Matrix of World Happiness Report Data Set

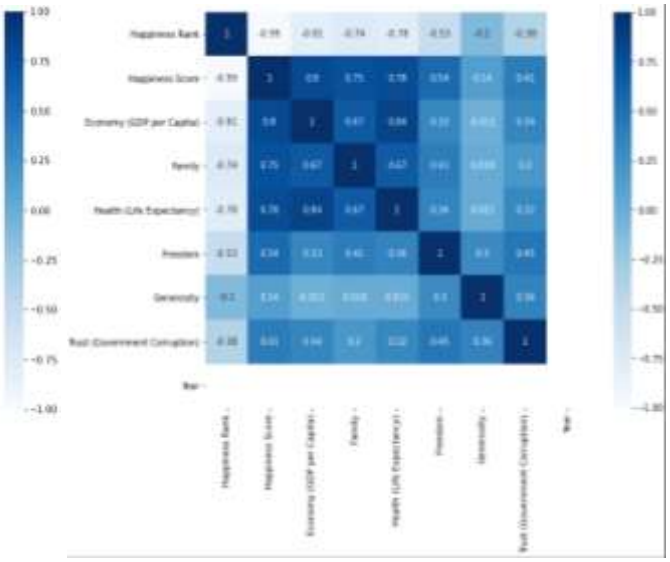
Left: for 2015 *Right: for 2016*



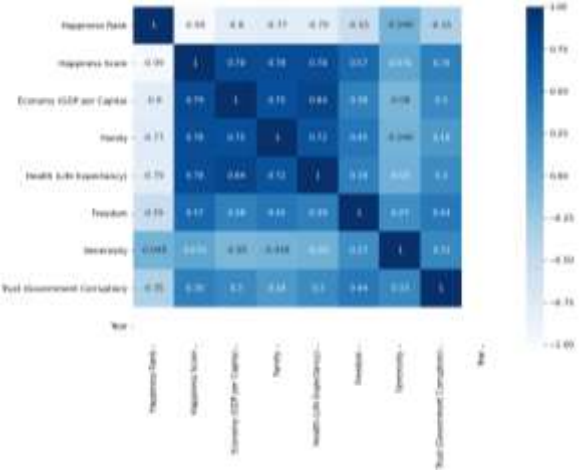


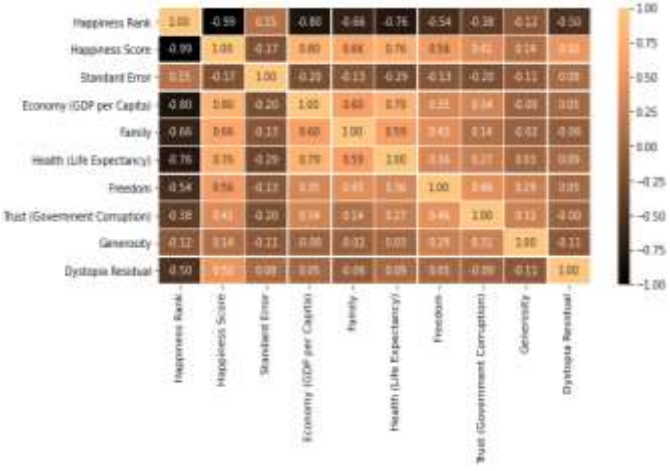
Left: for 2017 *Right: for 2018*





Left: for 2019 *Right: for 2015 -2019*





The above heat maps are showing how is the correlation between each variable, this report will concentrate the highly correlated variables. Happiness rank and happiness score have negative correlation, while happiness score is increasing, ranking is going to decrease. The correlation analysis shows that the main variables associated with the happiness score are GDP per capita, Life Expectancy and Social Support. It is because happiness scores are highly correlated with Economy(GDP), and somewhat positive related with Health(Life Expectations), Family Satisfaction, and Freedom also but has very low relation with Trust in Government in average case, which makes the situation very critical. On the other hand, Generosity is not related with the score in this analysis. This feature has less than 0.5 correlation which is very bad. GDP is the main factor which is effecting others as Family, Life Expectations and Freedom. However, we will need further investigation.

3.3 Principal Components Analysis (PCA)

In machine learning, sometimes too much data can be a bad thing. More features or dimensions can decrease a model's accuracy since there is more data that needs to be generalized.

The idea here is to use the dimensionality reduction technique that is a really powerful way to deal with huge datasets. The main purpose is to reduce the complexity of a model and avoid overfitting. There are two main categories of dimensionality reduction: feature selection and feature extraction. Feature selection is only keeping the most relevant variables from the original dataset, whereas feature extraction constructs a new set of feature properties based on a combination of the old ones.

This section will explore feature extraction. In practice, feature extraction is not only used to improve storage space, to store more information, or the computational efficiency of the learning algorithm, but can also keep or improve the predictive model's performance by reducing the dimensionality.

Specifically, the Principal Component Analysis(PCA) algorithm, will be discussed here. It is widely used across different fields, most prominently for feature extraction and dimensionality reduction. In a nutshell, PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.

Since the PCA directions are highly sensitive to data scaling, it needs to standardize the features prior to conduct PCA if the features were measured on different scales, and also ensures equal importance to all features. Let's perform a PCA step by step:

- Separate the predicted X variables and the target Y variable for each dataset:
    - Y variable for Retails: Weekly sales

    - Y variable for Diabetes: Outcome
    - Y variable for World Happiness Report: Happiness score
- Apply the function of sklearn's StandardScaler to standardize X features before performing the PCA transformation.
- Calculate standard deviation and cumulative explained variance across all PCs that of the PCA object within a data frame.
- Find the variance in each data set can be explained by the numbers of principal components in the graph
- Return PCA with components that explain 80% of the variance
- Create a new data frame with these principal components and concatenate it with the column of Y variables
- Generate plots with the reduced principal component with each dataset

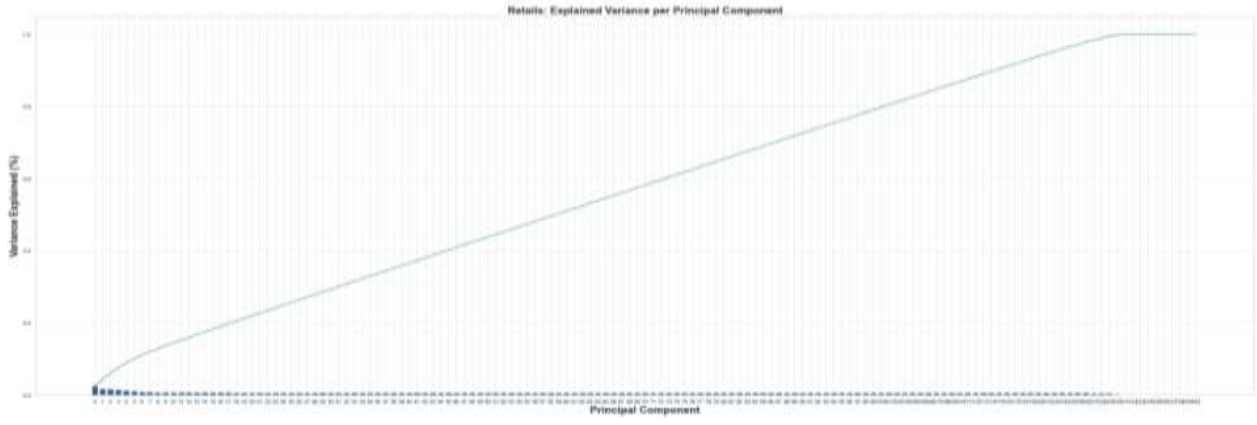
Let's look at the results of each dataset:

Table 7: Principal Component Analysis of Retails Data Set

	pca1	pca2	pca3	pca4	pca5	pca6	pca7	pca8	pca9	pca10	...	pca132	pca133	pca134	pca135	pca136	pca137	pca138	pca139	pca140	pca141
Standard deviation	1.9032	1.5933	1.5327	1.5044	1.3470	1.3115	1.1768	1.1214	1.1040	1.0786	...	0.4397	0.1927	0.0351	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Proportion of variance	0.0257	0.0180	0.0167	0.0161	0.0129	0.0122	0.0098	0.0089	0.0086	0.0083	...	0.0014	0.0003	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Cumulative proportion	0.0257	0.0437	0.0604	0.0764	0.0893	0.1015	0.1113	0.1202	0.1289	0.1371	...	0.9997	1.0000	1.0000	1.0	1.0	1.0	1.0	1.0	1.0	1.0

	pca1	pca2	pca3	pca4	pca5	pca6	pca7	pca8	pca9	pca10	...	pca93	pca94	pca95	pca96	pca97	pca98	pca99	pca100	pca101	Weekly_Sales
0	0.766687	-0.401285	-0.824860	0.051637	0.335047	-0.240079	-0.907099	0.676293	-0.212929	0.289520	...	0.166725	-0.292622	-0.142597	-0.680312	-0.927755	0.891267	-0.284706	-0.899653	0.462257	24924.50
1	0.770668	0.183070	-0.803242	3.248048	1.207318	-0.250334	-0.000587	1.040572	-0.230055	0.226438	...	0.145103	-0.324979	-0.175483	-0.984912	-1.040288	0.959658	-0.291628	-0.716230	0.354687	46038.49
2	0.763400	-0.381970	-0.835416	0.078374	0.329070	-0.217527	-1.003058	0.888733	-0.190756	0.300691	...	0.194795	-0.291528	-0.147296	-0.888892	-0.926230	0.894767	-0.287077	-0.886741	0.466735	41595.55
3	0.759706	-0.437368	-0.838663	0.023114	0.371530	-0.270971	-0.821882	0.879785	-0.180762	0.314292	...	0.201815	-0.294310	-0.144360	-0.894295	-0.927755	0.891471	-0.286006	-0.899737	0.461280	19483.54
4	0.757854	-0.436268	-0.823429	0.012853	0.358437	-0.260129	-0.777636	0.864060	-0.222556	0.288550	...	0.203865	-0.294367	-0.138581	-0.879373	-0.926258	0.887859	-0.282945	-0.883587	0.385289	21927.80
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
421565	-0.886605	0.437814	-0.990462	-0.611603	0.109273	-0.766598	0.126999	-0.681567	-0.644292	-0.717465	...	-0.188486	-1.115448	0.417700	1.626537	0.939901	0.426866	-0.686178	0.226179	-0.762283	880.37
421566	-0.857577	0.479053	-0.964854	-0.524153	0.189618	-0.653846	0.069682	-0.649699	-0.963339	-0.613330	...	-0.138504	-1.056723	0.427362	1.623633	0.989108	0.382002	-0.638498	0.268044	-0.754550	926.10
421567	-0.891617	0.490263	-0.131297	-0.386269	-0.175614	-0.904878	-0.194220	-0.846094	-0.607045	-0.674393	...	-0.147985	-1.055188	0.415130	1.636194	0.986221	0.347857	-0.616593	0.246039	-0.732278	1061.82
421568	-0.939621	0.451181	-0.153839	-0.359989	-0.247211	-0.861439	-0.147388	-0.785820	-1.034476	-0.640917	...	-0.137362	-1.056601	0.423807	1.633341	0.978418	0.338646	-0.693092	0.248380	-0.723040	780.91
421569	-0.933688	0.434843	-0.160567	-0.475629	-0.150755	-0.910980	-0.062899	-0.676020	-1.040956	-0.519716	...	-0.161254	-1.108415	0.418181	1.629515	0.963598	0.403428	-0.647008	0.223973	-0.751886	1076.80

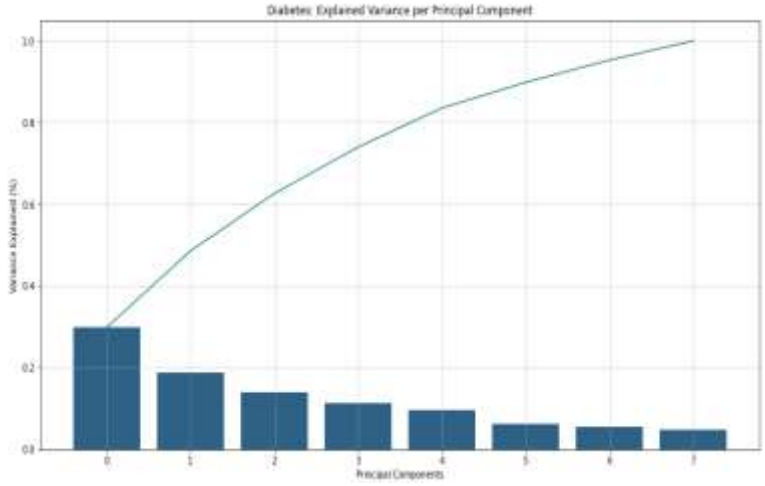
Figure 9: Principal Component Plot of Retails Data Set



The table 7 displays the Retail dataset has 141 total principal components, but from figure 9 above, 101 principal components help explain 80% of the variation, so this report will be using these 101 principal components in the model selection in the Retails dataset.

Table 8: Principal Component Analysis of Diabetes Data Set

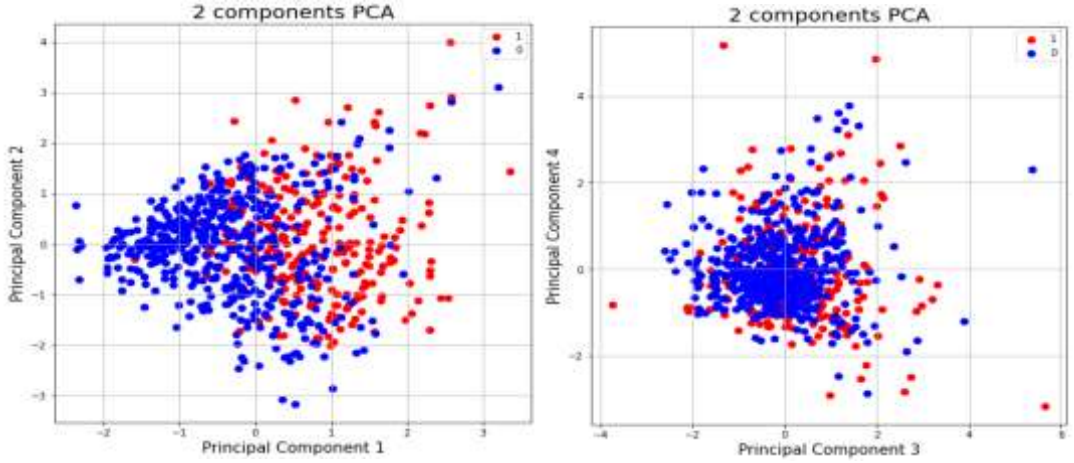
	pca1	pca2	pca3	pca4	pca5	pca6	pca7	pca8
Standard deviation	1.5455	1.2256	1.0604	0.9550	0.8747	0.7110	0.6604	0.6137
Proportion of variance	0.2982	0.1875	0.1404	0.1138	0.0955	0.0631	0.0544	0.0470
Cumulative proportion	0.2982	0.4857	0.6261	0.7399	0.8354	0.8985	0.9530	1.0000

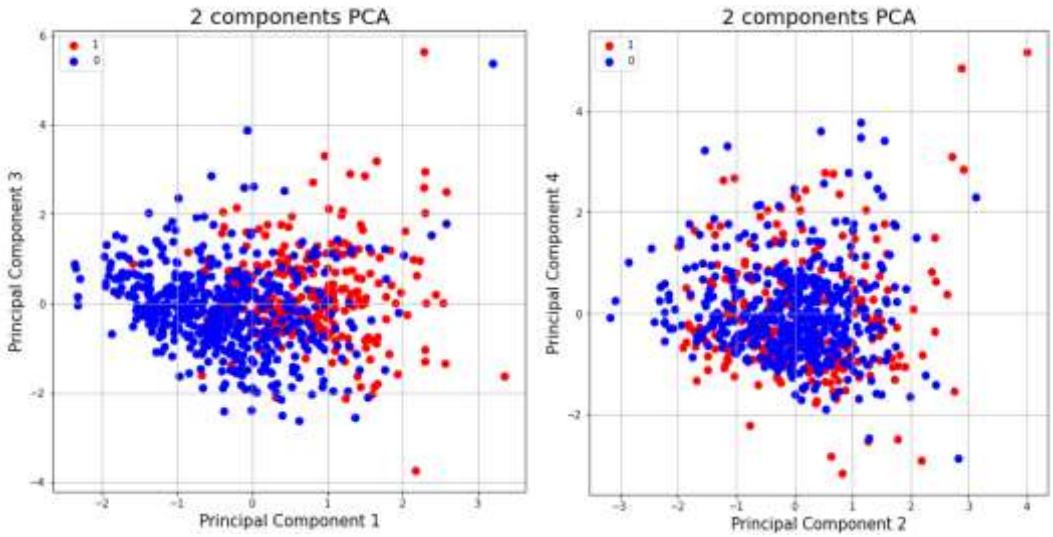


	pca1	pca2	pca3	pca4	Outcome
0	1.019595	-0.500532	0.417502	0.433301	1.0
1	-1.019699	0.064977	-0.321451	0.001115	0.0
2	0.475286	-0.443413	1.722515	0.108900	1.0
3	-1.378927	0.218064	-0.354564	-0.662579	0.0
4	0.520684	2.861177	1.960060	4.858117	1.0
...	...	...	...	...	...
763	1.377576	-1.646082	-1.120050	-0.059504	0.0
764	-0.303980	0.495480	-0.439911	-0.383012	0.0
765	-0.578845	-0.682969	0.216956	-0.531610	0.0
766	-0.045925	-0.003569	0.529178	-0.371287	1.0
767	-0.836363	0.527581	-0.688171	-0.287865	0.0

As the table 8 shown, the Diabetes dataset has 8 total principal components, but this report will be using 4 principal components in the model selection as they help explain 80% of the variation in this dataset. This report also plots the relationship between 4 components.

Figure 10: Principal Component Scatter Plot of Diabetes Data Set

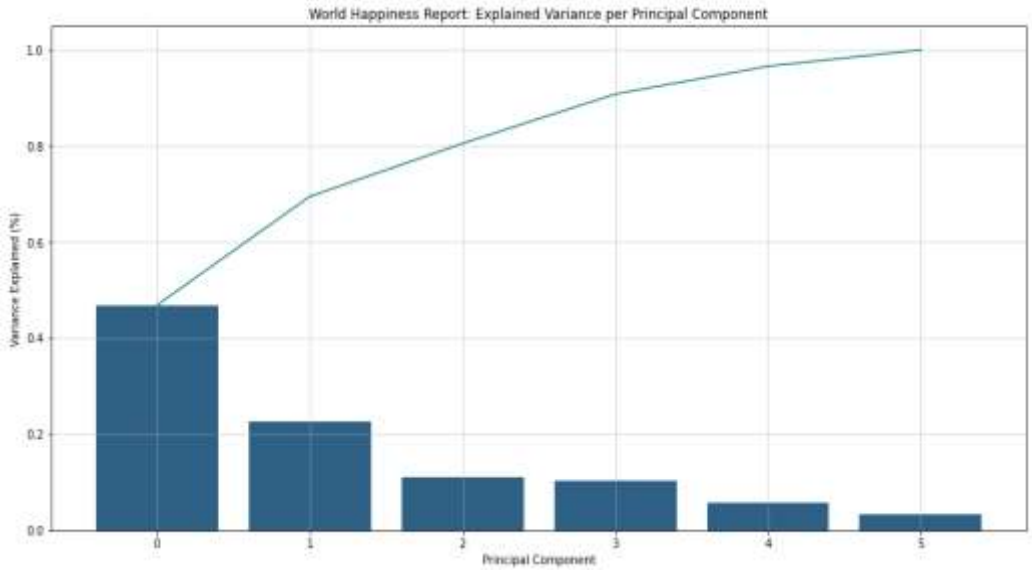




In the figure 10, we can see that the component 1 and component 2 are separated, also same for component 1 and 3, but the component 2 and component 3 tend to be overlapping.

Table 9: Principal Component Analysis of World Happiness Report Data Set

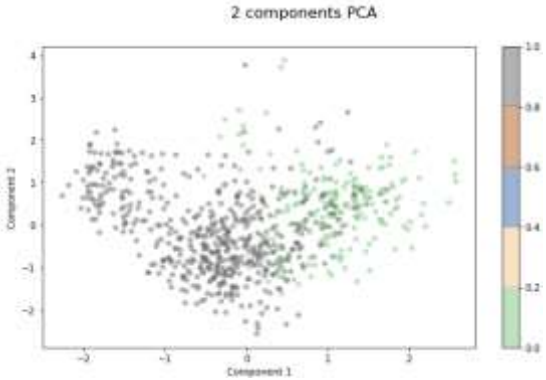
	pca1	pca2	pca3	pca4	pca5	pca6
Standard deviation	1.6777	1.1663	0.8142	0.7859	0.5908	0.4512
Proportion of variance	0.4685	0.2264	0.1103	0.1028	0.0581	0.0339
Cumulative proportion	0.4685	0.6949	0.8052	0.9080	0.9661	1.0000



Similarly, the World Happiness Report dataset has 6 total principal components as table 9 shown.

Figure 11: Principal Component Plot of World Happiness Report Data Set

	pca1	pca2	Happiness Score	New Happiness Score Name	New Happiness Score
1	-1.938449	1.225627	7.587	High Happiness	1
2	-1.436071	0.752206	7.561	High Happiness	1
3	-1.932981	1.737262	7.527	High Happiness	1
4	-1.838212	1.290335	7.522	High Happiness	1
5	-1.701686	1.668422	7.427	High Happiness	1
...	...	...	...	...	...
723	-0.040672	1.877183	3.334	Low happyiness	0
724	0.535367	0.800211	3.231	Low happyiness	0
725	2.036727	-0.546017	3.203	Low happyiness	0
726	2.576366	1.027719	3.083	Low happyiness	0
727	1.938832	-0.029029	2.853	Low happyiness	0



Moving forward, this report will be using 2 principal components in the model selection as they help explain 80% of the variation in the World Happiness Report dataset. In the right figure, the gray dot shows the high happiness score, and the green one is the low happiness score.

Therefore, PCA helps to identify patterns in data based on the correlation between features, and the goal is to maintain most of the relevant information. Let's move to the modeling.

4 Modeling Process

This section builds popular supervised learning algorithms, including various regression and classification models, and compared to each other based on their predictive score on the hold-out samples. The dataset contains the predicted X and target y values that building up modeling by comparing both the original datasets and after using PCA techniques. It splits train and test into an 80:20. This modeling uses cross validation technique K folds or through shuffle split for parameter tuning to identify the best model.

4.1 Regression Modeling – Retails

This dataset mainly focuses on X regression models due to the running time limit, including Linear regression, Ridge regression, Lasso regression, K Nearest Neighbor, and Random Forest. About 140 predicted features identified are "Temperature", "Fuel Price", "MarkDown1-5", "CPI", "Unemployment", "Size", "Not Holiday", "Holiday", "A", "B", "C", "Store1-45" and "Dept1-99". This section also uses the cross-validation technique to predict the response on the test data, but tree models use 5-fold here, so we cannot compare them with other models. Hence, this section is more focused on the r squared comparison.

4.1.1 Scores of Best Model

The model comparison is evaluated by the following metrics as the table showing below:

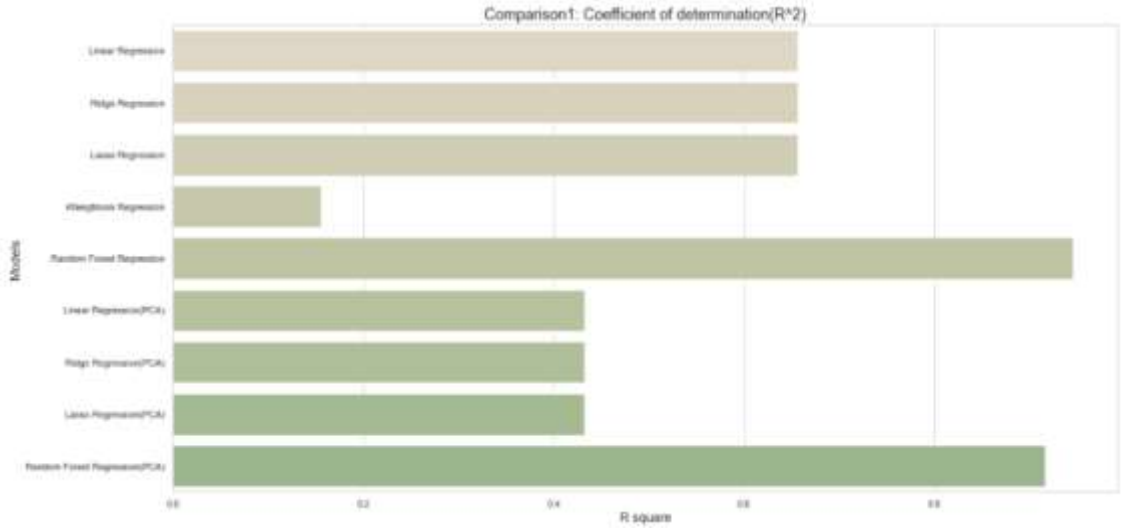
Table 10: Model Comparison of Retails Data Set

	Models	Full sample R-squared	Train Score	Test Score	Test Score(CV)	Coefficient of determination(R^2)	Test MSE	Test ME	Test RMSE	Test MAE	Time
1	Linear Regression	0.657876	0.658101	0.656986	0.650838	0.656411	1.788718e+08	31.5464	13374.2966	8164.9902	18.855679
2	Ridge Regression	0.657793	0.658047	0.656925	0.667297	0.656320	1.789197e+08	31.5572	13376.0855	8156.8053	43.387972
3	Lasso Regression	0.657876	0.658101	0.656986	0.661921	0.656412	1.788716e+08	31.5622	13374.2887	8164.8695	2604.687244
4	KNeighbors Regression	0.402077	0.464691	0.155563	0.16294	0.155563	4.403493e+08	-13.5805	20984.5023	13012.4208	49379.338378
5	Random Forest Regression	0.982988	0.651016	0.523995	N/A	0.945594	2.837111e+07	-26.8338	5326.4541	1890.7518	23222.266643
6	Linear Regression(PCA)	0.433005	0.432924	0.433326	0.430156	0.432646	2.955041e+08	84.7301	17190.2332	11577.7605	13.855897
7	Ridge Regression(PCA)	0.433005	0.432924	0.433326	0.436785	0.432646	2.955044e+08	84.7246	17190.2417	11577.5911	31.184158
8	Lasso Regression(PCA)	0.433005	0.432924	0.433326	0.439044	0.432646	2.955041e+08	84.7298	17190.2337	11577.7549	40.544677
9	Random Forest Regression(PCA)	0.974369	0.925851	0.516962	N/A	0.916205	4.369641e+07	-154.3256	6610.3257	2292.8656	50809.382582

	Models	Train Set Mean Score	Test Set Mean Score
1	Linear Regression	0.657597	0.658999
2	Ridge Regression	0.656736	0.662235
3	Lasso Regression	0.657462	0.659567
4	KNeighbors Regression	1.000000	-0.106371
5	Random Forest Regression	0.642067	0.628485
6	Linear Regression(PCA)	0.433576	0.430747
7	Ridge Regression(PCA)	0.432800	0.433775
8	Lasso Regression(PCA)	0.432892	0.433227
9	Random Forest Regression(PCA)	0.916647	0.887233

As Table 10 shown, Radom Forest has the highest test score than linear, ridge, lasso and KNN. After using PCA techniques, Radom Forest also has the best score of 88.7% with the best parameter chosen. Linear, Ridge and Lasso perform similarly, but KNN does not fit well.

Figure 12: Model Comparison Bar Plot of Retails Data Set



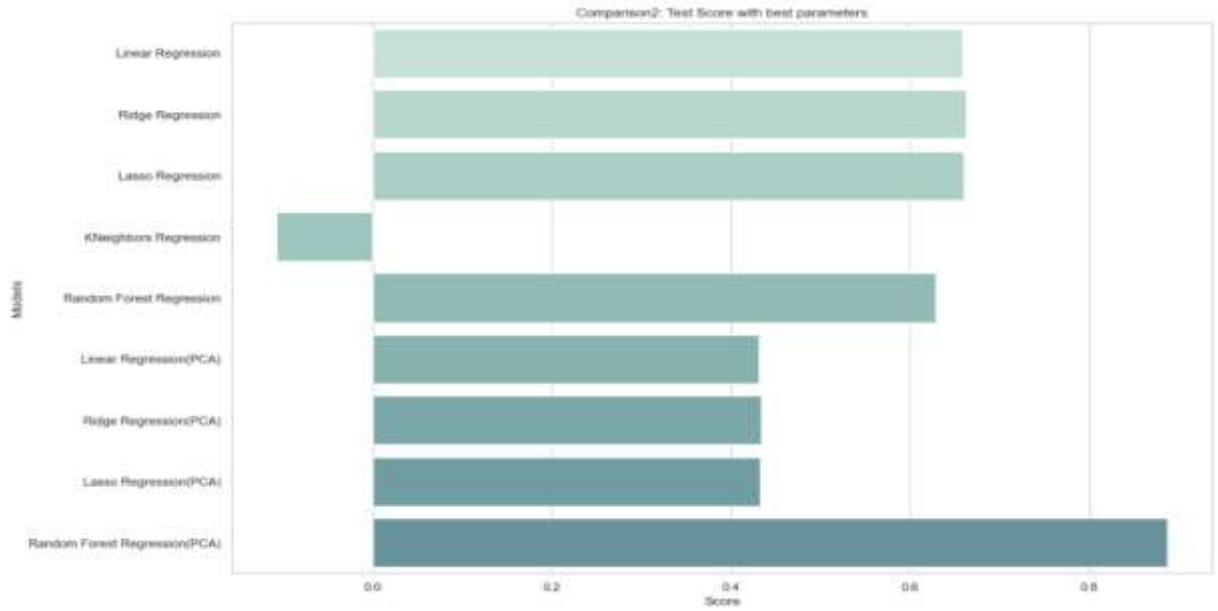


Figure 12 above visualize the distribution of all the models based on the best parameters chosen for each model. Here, Random Forest Regression is considered as the best model among all models. KNN has a negative result, so this report does not consider it to use the PCA technique.

4.1.2 Overfitting Diagnosis

It is an important part to avoid overfitting in the modeling. For each model, we have already reduced 141 to 101 dimensions by using the principal component technique to prevent data overfitting. Another good way did in this modeling process is to add some restrictions like pruning the tree model for limiting the numbers of nodes. This report also adjusts the alpha and k in models.

4.1.3 Unseen Data

This section generates 200 total samples of unseen data from the original retail dataset. The unseen data separated into two parts: predictors who had a positive correlation that is temperature, Markdown1-3, Markdown5, Not Holiday, Holiday, and Size, and those who had a negative correlation that are Fuel Price, MarkDown4, CPI, and Unemployment with Weekly Sales. This figure shows two groups of results as below:

Figure 13: Test Unseen Data Result of Retails Data Set

```
------------------------------
Random Forest Regression(Best Model):
------------------------------
Positive coefficient upward
MSE: 249309668
R2 Score: 0.6278
RMSE: 15789.543

Negative coefficient downward
MSE: 103961160
R2 Score: 0.6926
RMSE: 10196.1346
```

4.1.4 Results Interpretation

For testing the unseen data, we use the random forest regression model identified as the best model obtained in the 4.1.1 section. The test result for positive coefficient upward is 62.78%, and for negative coefficient downward is 69.26%.

4.2 Classification Modeling – Diabetes

This dataset mainly applies 6 classification models, including K Nearest Neighbor, Gaussian Naive Bayes, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. 8 predicted features identified are “Pregnancies”, “Glucose”, “Blood Pressure”, “Skin Thickness”, “Insulin”, “BMI”, “Diabetes Pedigree Function”, and “Age”. This section uses the cross-validation technique to predict the response on the test data and produce a confusion matrix comparing the test labels to the predicted test labels.

4.2.1 Scores of Best Model

We evaluate the performance of the model using the following metrics and the score comparison is described as the table showing below:

Table 11: Model Comparison of Diabetes Data Set

	Models	Train Score	Test Score	Test Score(CV)	Precision(0)	Precision(1)	Recall(0)	Recall(1)	ME	RMSE	MAE	Time
1	KNeighbors Classifier	0.921824	0.863636	0.839675	0.903226	0.754098	0.848485	0.836364	-0.0390	0.3948	0.1558	11.289974
2	GaussianNB	0.767101	0.766234	0.764442	0.815534	0.705882	0.848485	0.654545	0.0260	0.4699	0.2208	4.014194
3	Logistic Regression	0.763844	0.779221	0.768532	0.847619	0.795918	0.898990	0.709091	0.0390	0.4109	0.1688	6.677235
4	Decision Tree Classifier	1.000000	0.837662	0.843649	0.800000	0.693878	0.848485	0.618182	0.0390	0.4835	0.2338	5.111646
5	Random Forest Classifier	1.000000	0.883117	0.879883	0.861386	0.773585	0.878788	0.878788	0.0130	0.4109	0.1688	133.940368
6	Gradient Boosting Classifier	0.993485	0.870130	0.880610	0.860000	0.759259	0.868687	0.745455	0.0065	0.4187	0.1753	96.705381
7	KNeighbors Classifier(PCA)	0.814332	0.824675	0.781597	0.697842	0.866667	0.979798	0.236364	0.2597	0.5345	0.2857	11.161147
8	GaussianNB(PCA)	0.760586	0.759740	0.758935	0.801887	0.708333	0.858586	0.618182	0.0455	0.4767	0.2273	4.007833
9	Logistic Regression(PCA)	0.775244	0.766234	0.777195	0.839623	0.791667	0.898990	0.690909	0.0455	0.4187	0.1753	3.913381
10	Decision Tree Classifier(PCA)	1.000000	0.759740	0.737273	0.815217	0.612903	0.757576	0.690909	-0.0455	0.5160	0.2662	4.867533
11	Random Forest Classifier(PCA)	1.000000	0.792208	0.780532	0.816327	0.660714	0.808081	0.808081	-0.0065	0.4902	0.2403	139.669056
12	Gradient Boosting Classifier(PCA)	0.933225	0.805195	0.788078	0.831683	0.716981	0.848485	0.690909	0.0130	0.4558	0.2078	87.157983

Note that confusion matrix for each model attached in the relevant coding file.

As Table 11 shown, Radom Forest and Gradient Boosting have really close test scores, which are higher than other original models. After using PCA techniques, Gradient Boosting has the best score of 80.5%. Another attractive model is K Nearest Neighbor that also has a higher score before and after using PCA techniques, 86.4% and 82.5%, which is in the third place.

Figure 14: Model Comparison Bar Plot of Diabetes Data Set

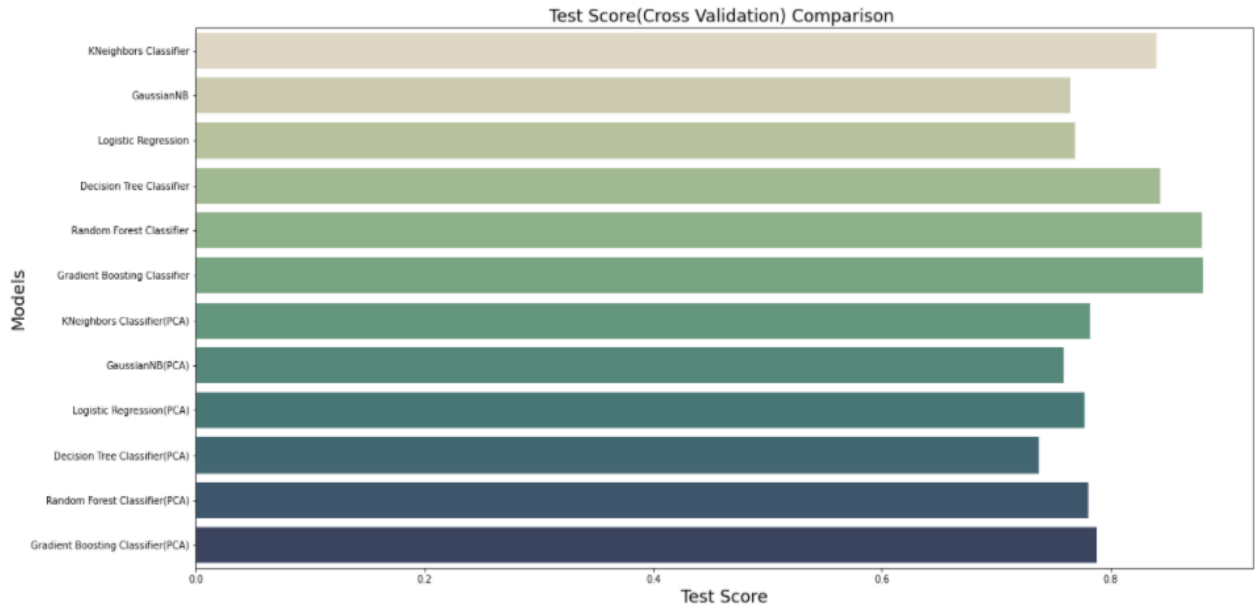


Figure 14 above visualize the distribution of all the models based on using the cross-validation technique. Hence, Gradient Boosting Classification is considered as the best model among all models.

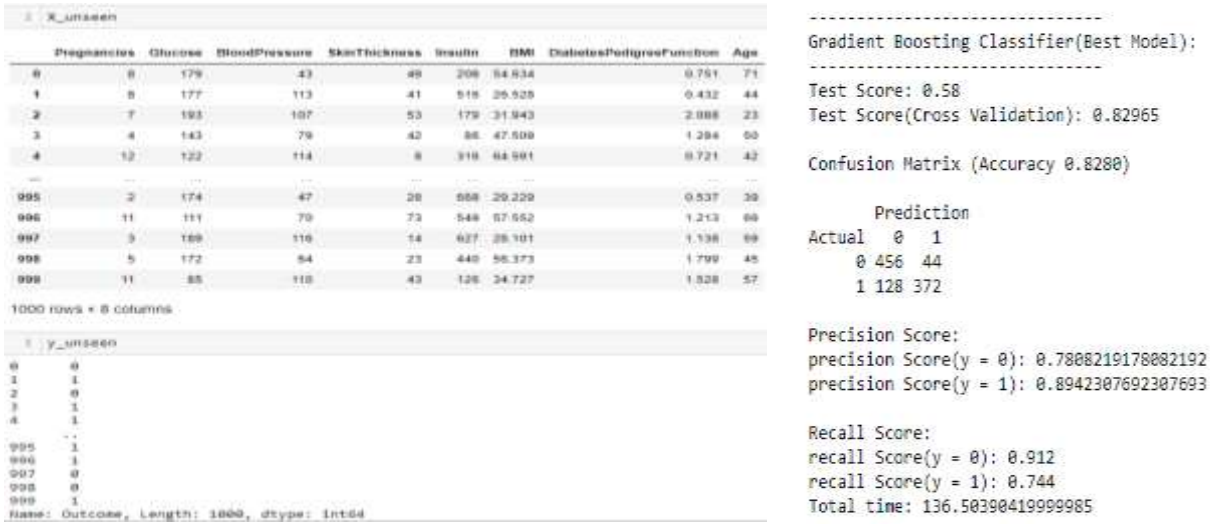
4.2.2 Overfitting Diagnosis

This section was careful to avoid overfitting and other negative influences by using the information gain ratio in the modeling. We have already reduced 8 to 4 dimensions for each model by using the principal component technique to prevent data overfitting. Another effective way in this modeling process is to add some restrictions like pruning the tree models for limiting the numbers of nodes. This report also adjusts the alpha and k in models.

4.2.3 Unseen Data

The unseen dataset is collected based on the y variable Outcome by separating the results if diagnosed with Diabetes(classified as 0 or 1). This section generates 500 random data and use predicted features as same as in the 4.2 section.

Figure 15: Test Unseen Data Result of Diabetes Data Set



```
X_unseen
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0	179	43	49	208	54.634	0.751	71
1	8	177	113	41	518	26.528	0.432	44
2	7	193	107	53	179	31.943	2.088	23
3	4	143	79	42	86	47.509	1.294	60
4	12	122	114	8	318	64.991	0.721	42
...	...	...	...	...	...	...	...	...
995	2	174	47	28	568	29.229	0.537	39
996	11	111	70	73	549	57.552	1.213	66
997	3	189	116	14	627	28.101	1.138	69
998	5	172	64	23	440	56.373	1.799	45
999	11	85	110	43	126	34.727	1.328	57

1000 rows × 8 columns

```
y_unseen
```

```
0      0
1      1
2      0
3      1
4      1
      ..
995    1
996    1
997    0
998    0
999    1
Name: Outcome, Length: 1000, dtype: int64
```

```
--------------------------------
Gradient Boosting Classifier(Best Model):
--------------------------------
Test Score: 0.58
Test Score(Cross Validation): 0.82965

Confusion Matrix (Accuracy 0.8280)

       Prediction
Actual   0   1
     0 456  44
     1 128 372

Precision Score:
precision Score(y = 0): 0.7808219178082192
precision Score(y = 1): 0.8942307692307693

Recall Score:
recall Score(y = 0): 0.912
recall Score(y = 1): 0.744
Total time: 136.50390419999985
```

4.2.4 Results Interpretation

For testing the unseen data, we use the gradient boosting classification model identified as the best model obtained in the 4.2.1 section. It has a good result with the cross-validation score of 82.97%.

4.3 Triage Modeling – World Happiness Report

This dataset mainly focuses on 7 regression models, including Linear regression, Ridge regression, Lasso regression, K Nearest Neighbor, Decision Tree, Random Forest, and Gradient Boosting. 6 predicted features identified are “Economy (GDP per Capita)”, “Family”, “Health (Life Expectancy)”, “Freedom”, “Trust (Government Corruption)”, and “Generosity”. This section also uses the cross-validation technique to predict the response on the test data, but tree models use 5-fold here, so we cannot compare them with other models. Hence, this section is more focused on the r squared comparison.

4.3.1 Scores of Best Model

The model performance is described as the table showing below and here are the metrics we used:

Table 12: Model Comparison of World Happiness Report Data Set

	Models	Full sample R-squared	Train Score	Test Score	Test Score(CV)	Coefficient of determination(R^2)	Test MSE	Test ME	Test RMSE	Test MAE	Time
1	Linear Regression	0.772234	0.778829	0.744303	0.760822	0.744303	0.328205	-0.0173	0.5729	0.4395	0.958174
2	Ridge Regression	0.763160	0.777730	0.747247	0.813876	0.743494	0.329243	-0.0162	0.5738	0.4532	82.712537
3	Lasso Regression	0.763871	0.770641	0.742471	0.818233	0.741482	0.331825	-0.0218	0.5760	0.4458	78.069637
4	KNeighbors Regression	0.755246	0.793708	0.775150	0.809801	0.744619	0.327799	-0.0539	0.5725	0.4578	136.088674
5	Decision Tree Regression	0.913275	0.785791	0.695100	N/A	0.546522	0.582070	0.0128	0.7629	0.5819	1.476493
6	Random Forest Regression	0.939870	0.974508	0.799087	N/A	0.791463	0.267671	0.0320	0.5174	0.4118	601.353940
7	Gradient Boosting Regression	0.894372	1.000000	0.799647	N/A	0.771642	0.293113	0.0394	0.5414	0.4274	2046.696755
8	Linear Regression(PCA)	0.763828	0.772153	0.728584	0.758624	0.728584	0.348381	-0.0036	0.5902	0.4574	1.727205
9	Ridge Regression(PCA)	0.763447	0.772061	0.727862	0.824723	0.727451	0.349835	-0.0038	0.5915	0.4608	68.727916
10	Lasso Regression(PCA)	0.763333	0.772068	0.727073	0.830376	0.726593	0.350937	-0.0039	0.5924	0.4607	69.453456
11	KNeighbors Regression(PCA)	0.744297	0.752715	0.714322	0.810403	0.713242	0.368073	-0.0011	0.6067	0.4885	75.162484
12	Decision Tree Regression(PCA)	0.904865	0.807445	0.745720	N/A	0.502546	0.638516	0.0031	0.7991	0.6047	0.691040
13	Random Forest Regression(PCA)	0.915132	0.822509	0.765588	N/A	0.695798	0.390465	0.0156	0.6249	0.4853	249.018699
14	Gradient Boosting Regression(PCA)	0.852730	0.847275	0.761494	N/A	0.739267	0.334669	0.0085	0.5785	0.4611	1179.297946

	Models	Train Set Mean Score	Test Set Mean Score
1	Linear Regression	0.774361	0.760054
2	Ridge Regression	0.772967	0.762101
3	Lasso Regression	0.765616	0.755461
4	KNeighbors Regression	0.792480	0.777354
5	Decision Tree Regression	0.784143	0.699342
6	Random Forest Regression	0.974697	0.812880
7	Gradient Boosting Regression	1.000000	0.810562
8	Linear Regression(PCA)	0.764186	0.759289
9	Ridge Regression(PCA)	0.764661	0.756776
10	Lasso Regression(PCA)	0.763750	0.760931
11	KNeighbors Regression(PCA)	0.747600	0.736903
12	Decision Tree Regression(PCA)	0.805996	0.753810
13	Random Forest Regression(PCA)	0.821405	0.773918
14	Gradient Boosting Regression(PCA)	0.846806	0.770484

As Table 12 shown, Radom Forest has the best score with 79.9%, and the following is Gradient Boosting which has really close test scores with random forest. Their r squared is also higher than other original models, which are 79.15% and 77.16%. After using PCA techniques, Gradient Boosting has the best r squared of 73.93%. The random forest doesn't have a higher r squared. Another model is K Nearest Neighbor, which has higher r squared before and after using PCA techniques, 74.46%, and 71.32%.

Figure 16: Model Comparison Bar Plots of World Happiness Report Data Set

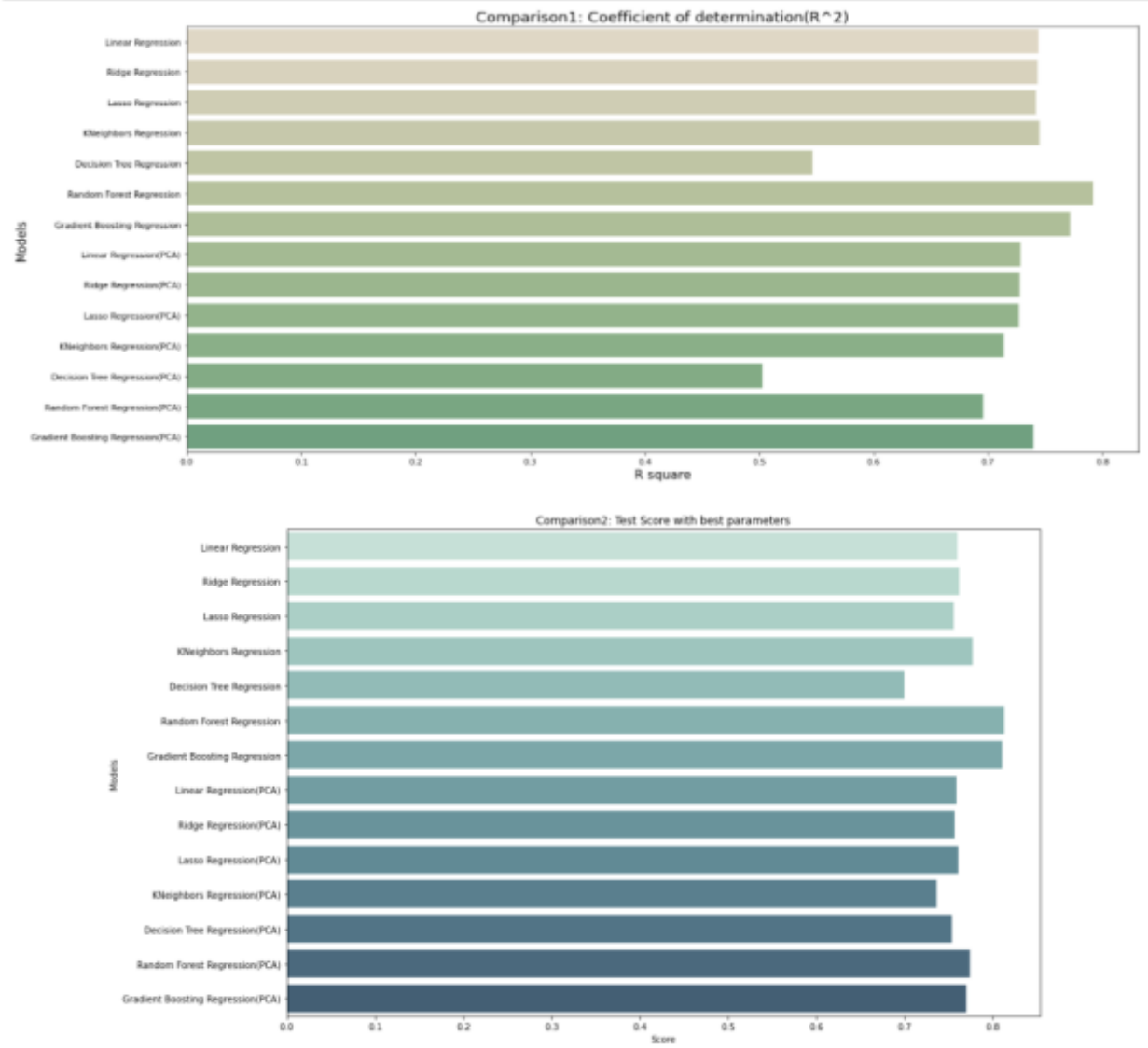


Figure 16 above visualize the distribution of all the models based on the best parameters chosen for each model. Here, Gradient Boosting Regression is considered as the best model among all models.

4.3.2 Overfitting Diagnosis

It is a crucial step to avoid overfitting in the modeling. We have already reduced 6 to 2 dimensions for each model using the principal component technique to prevent data overfitting. Another good way in this modeling process is to add some restrictions like pruning the tree models for limiting the numbers of nodes. This report also adjusts the alpha and k in models.

4.3.3 Unseen Data

In this section, the unseen dataset is the 2020 World Happiness Report collected from Kaggle, published by Mathurin Aché, Sustainable Development Solutions Network[4]. During the data preparation, it renamed most of the features to match with the name with the combing datasets of 2015-2019. Missing values also have been checked, and it goes into the modeling process. This unseen data uses the predicted features as same as in the 4.3 section.

Figure 17(a): Test Unseen Data Result of World Happiness Report Data Set

	param_max_depth	param_max_features	param_n_estimators	param_learning_rate	mean_train_score	mean_test_score	rank_test_score
569	2	2	100	0.08	0.949124	0.788944	1
611	4	2	100	0.08	0.997962	0.786015	2
843	6	2	150	0.1	1.000000	0.785834	3
401	3	2	100	0.06	0.976248	0.785725	4
26	3	2	250	0.02	0.970179	0.784569	5
...	...	...	...	...	...	...	...
1111	15	6	250	0.25	1.000000	0.606247	1130
1109	15	6	150	0.25	1.000000	0.604130	1131
1131	20	6	200	0.25	1.000000	0.603424	1132
1132	20	6	250	0.25	1.000000	0.601718	1133
1133	20	6	300	0.25	1.000000	0.596787	1134

1134 rows × 7 columns

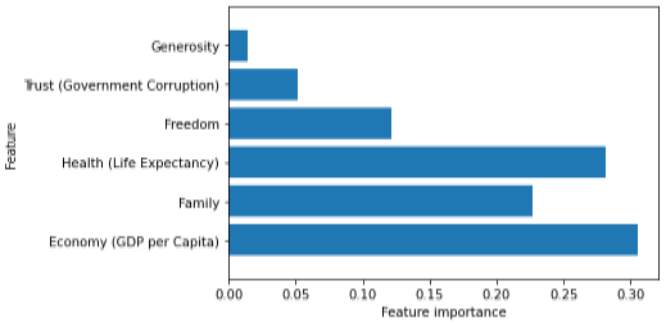
```
---- optimal parameters ----
{'learning_rate': 0.08, 'max_depth': 2, 'max_features': 2, 'n_estimators': 100}
---- best accuracy ----
0.7899444018536085
```

```
Mean squared error (MSE): 0.2764
Coefficient of determination (R^2): 0.6557

Train Set:

Regression statistics

                    Mean Error (ME) : -0.0000
     Root Mean Squared Error (RMSE) : 0.1212
          Mean Absolute Error (MAE) : 0.0927
        Mean Percentage Error (MPE) : -0.2122
Mean Absolute Percentage Error (MAPE) : 1.7953

Test Set:

Regression statistics

                    Mean Error (ME) : 0.1697
     Root Mean Squared Error (RMSE) : 0.5258
          Mean Absolute Error (MAE) : 0.4380
        Mean Percentage Error (MPE) : 2.8598
Mean Absolute Percentage Error (MAPE) : 7.8828
```

4.3.4 Results Interpretation

We use gradient boosting as the best model obtained in the 4.3.1 section to test the unseen data and check the predictive score. It has a good result with a score of 78.99%. And we found that the GDP is the most critical factor affecting happiness, and the following is Health(life expectancy).

Figure 17(b): Test Unseen Data Result of Feature Importance



[4] https://www.kaggle.com/mathurinache/world-happiness-report?select=2020.csv

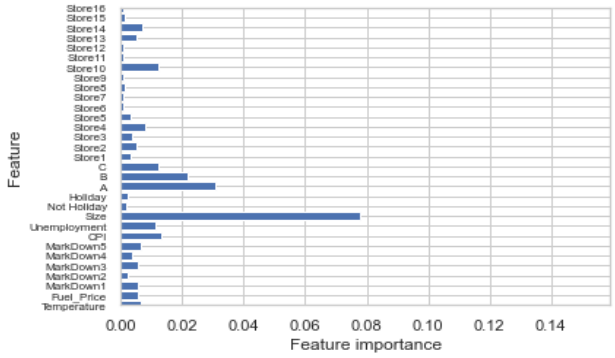
5 Conclusion and Suggestions

Through these three datasets, this report has discovered many relationships across 3 different groups of targets and their informative variables. Data mining-based approaches can be used to assess each group of predictor variables influencing each target, which are the trend of weekly sales, the risk of diabetes, and the cause of happiness scores. Some reflection and suggestions are provided below for Retails, Diabetes, and World Happiness Report.

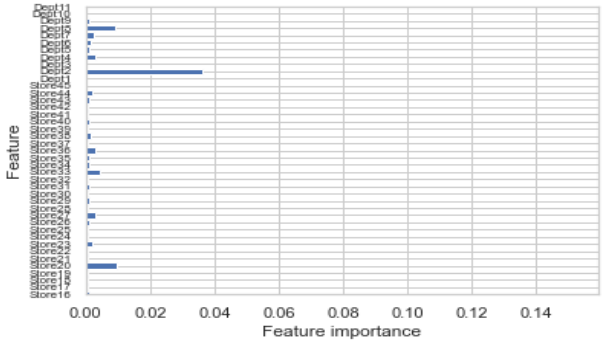
The retails dataset has been analyzed in detail above. This dataset was validated using modeling techniques employed that contain ridge, lasso, k nearest neighbor, and one tree models. These were built and evaluated to identify the best model to predict the weekly sales. From the r squared performance measures, the Random Forest Regression has concluded the best performing model. Figure 18 shows the most important factors such as Size, Dept13, Dept38, Dept40, Dept65, Dept 72, Dept 90, Dept92, and Dept 95. Since this dataset reduced 141 to 101 dimensions, so many other principal components were treated as important features, like pca1, pca47, pca48, pca49, and pca69, after using the data reduction technique.

Figure 18: Bar Plots Comparison of Feature Importance of Retails Data Set

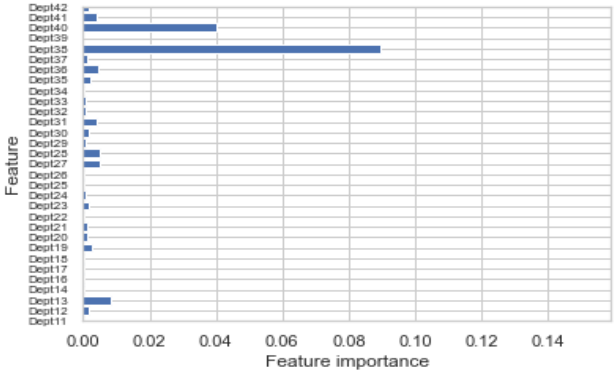
Variable 1 – 30



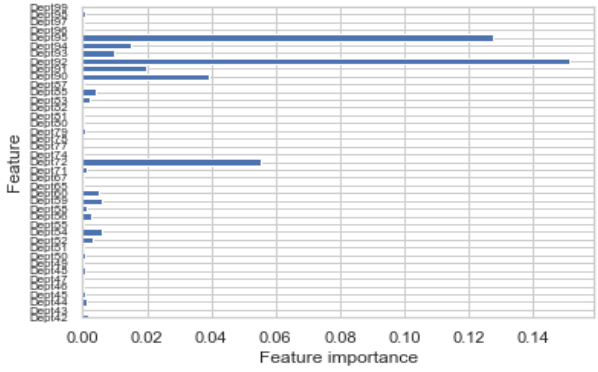
Variable 31 - 70



Variable 71 – 100

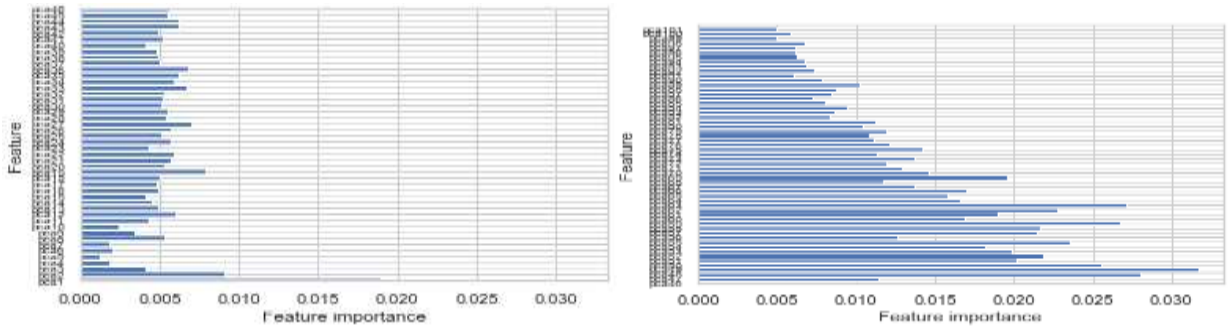


Variable 101 - 140



PCA Variable 1 – 45

PCA Variable 46 – 101

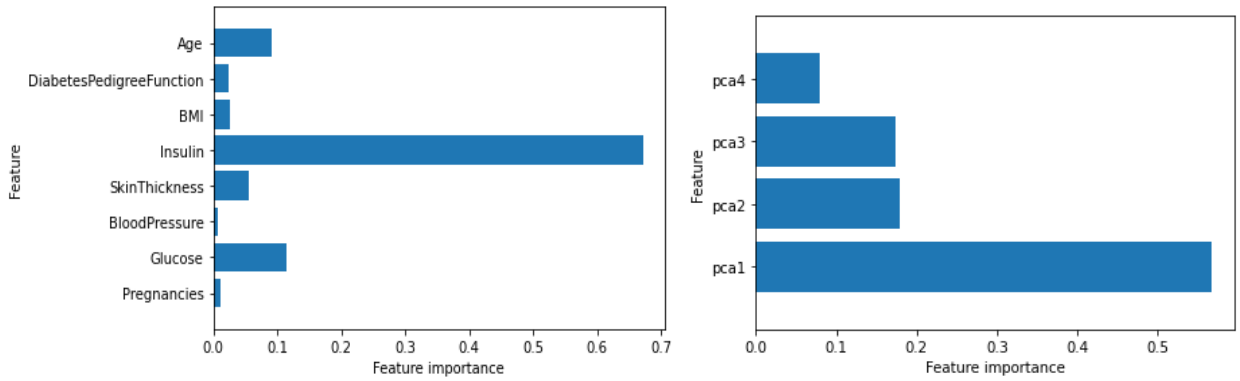


Note that this plot is of the best model chosen in the Section 4, other related plots in the relevant coding files.

We can see from the visualization of profit sales by departments that these six departments generate around a total of 31% sales, which are Dept38, Dept65, Dept40, Dept 72, Dept90, Dept92, and Dept 95 that each takes up 3.97%, 3.99%, 4.47%, 5.37%, 6.13%, and 6.61% of the most profits. This result corresponds to the important features as figure 18 shown below, therefore we predict these 6 departments will continue to generate sales for the next year. The suggestion is to give more attention on and improve the budgets for these 6 departments, also optimize resources from those departments that have loss sales.

The diabetes dataset has been explored in detail above. The patterns identified using data exploration methods were validated using modeling techniques employed. Classification models such as logistic regression, various tree models were built and evaluated to determine the best model to predict the occurrence of diabetes. From the cross-validation performance measures of sensitivity, the Gradient Boosting Classification has concluded the best performing model. In our findings based on Figure 19, the most important factors that caused diabetes are Insulin, Glucose, Age. And most features concentrated on the first principal component. Normally, blood glucose levels are tightly controlled by insulin. When the blood glucose elevates (for example, after eating food), insulin is released to normalize the glucose level. But in patients with diabetes, the absence or insufficient production of insulin causes hyperglycemia. Excessive glucose in the blood also resulting in diabetes, also damages kidneys, blood vessels, and skin increasing the possibility of a heart attack. Another significant factor is age because metabolism slows down as people get older. The decreased immunity of body cells due to their inability to absorb glucose, so elders are more vulnerable to infection and diabetes, further affects other vital body organs. Although it is a chronic disease, we should give more attention to diabetes. The suggestion is to avoid an irregular lifestyle, and it can be corrected with proper diet and exercise, like limit sugar intake and work out at least 30min to 1 hour everyday.

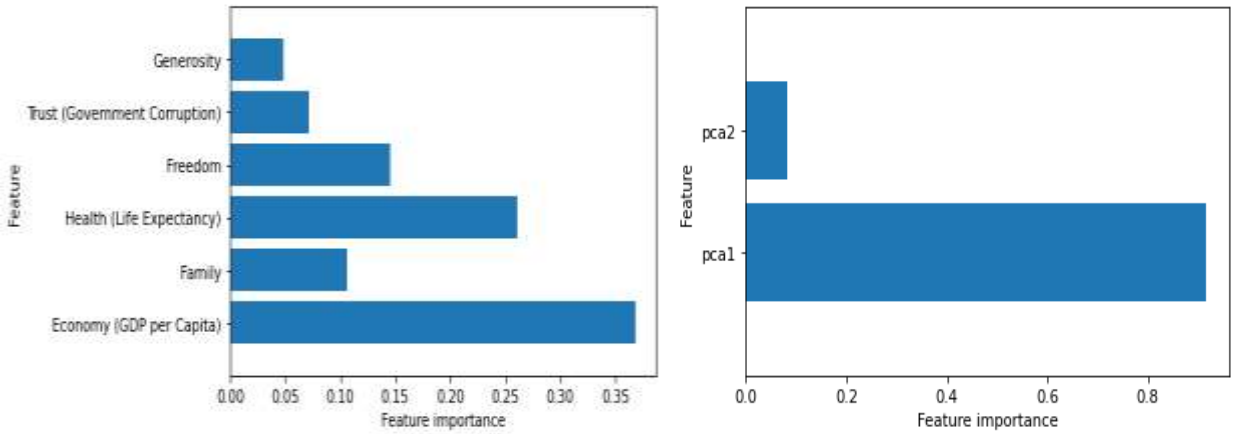
Figure 19: Bar Plot Comparison of Feature Importance of Diabetes Data Set



Note that this plot is of the best model chosen in the Section 4, other related plots in the relevant coding files.

The world happiness report dataset has also been examined in detail above. Data exploration methods were validated using modeling techniques employed that contain ridge, lasso, and various tree models. These were built and evaluated to identify the best model to predict the happiness score. From the r squared performance measures of sensitivity, the Gradient Boosting Regression has concluded the best performing model. Figure 20 showing that there is no single factor that can explain the happiness of people. Factors such as GDP per capita, health, and family all have the most important roles on happiness. We should consider all aspects together. Additionally, most features concentrated on the first principal component.

Figure 20: Bar Plot Comparison of Feature Importance of World Happiness Report Data Set



Note that this plot is of the best model chosen in the Section 4, other related plots in the relevant coding files.

We know very well that money does not buy happiness by itself but it provides the other factors to be happier as healthier life, trustable government, let family have higher life quality, freedom to make life choices, and freedom from corruption.

GDP is like a catalyzer which is affecting most of the factors. Developed countries are working on new technologies, such as AI, electrical cars, and most of them are ready for climate changes in next decades. They are also investing in clean energy, agricultural sciences, and cleaning the air from pollutants. They will be the best-survived countries against climate changes. So developed countries will easily keep their happiness and life standards in future. On the other hand, poor and underdeveloped countries will worsen day after day as they have limited resources, high and not educated population, uncontrollable and increased pollution, and wars, their future is very dark. Climate change will show the effects very fastly in near future. We will not see a better picture in the future with all these results and conditions. However, we still hope and suggest underdeveloped countries can improve their current limited circumstances and increase their happiness score while developed countries will be keeping their status and be happier.

To sum up, each target has several factors, Insulin, Size, GDP are the most powerful factors in each group, but not the only one. There are other potential issues, risks, and elements relating to the targets. For example, the environment factor affects the store location, for the place where diabetes patients live, and happiness with the living area's satisfaction. For future study, we could try to gather more information that are possibly influencing sales, diabetes, and happiness score to improve our prediction.

6 Reference

Berk, R.A. (2004). Regression analysis. A constructive critique. Thousand Oaks, CA: Sage.

Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. 2020. World Happiness Report 2020. New York: Sustainable Development Solutions Network.

Appendix Relevant Coding

Please see three attached Jupiter Notebook Files: Project 2-1, Project 2-2 and Project2-3.