Crash Course in Data Science

Matthew Motley and Ritesh Pachgade

5/10/2019

1 Introduction and Overview

1.1 General Tutorial Information

This tutorial is a crash course to data science and is meant for beginners. Users will be walked through the entire data science pipeline: data curation, parsing, and management; exploratory data analysis; hypothesis testing and machine learning to provide analysis" (Bravo 2019), using a real world data set example of which can be downloaded here: World Happiness Report -> https://www.kaggle.com/unsdsn/world-happiness (https://www.kaggle.com/unsdsn/world-happiness) An example problem will be specified and solved in this tutorial using this data set, while insights will be explained at the end of a section elaborating on what was learned in that section.

For data processing we will be using the R Data Science Toolbox, which is a combination of R and RStudio. Various packages will be used that can be installed directly in the RStudio. Data processing will be performed both R and SQL programming languages.

A tutorial for setting up the R Data Science Toolbox (R and RStudio) can be found here: Setting up R Data Science Toolbox -> http://www.hcbravo.org/IntroDataSci/bookdown-notes/setting-up-the-r-data-science-toolbox.html (http://www.hcbravo.org/IntroDataSci/bookdown-notes/setting-up-the-r-data-science-toolbox.html) (Bravo 2019).

External sources will be used in defining or presenting the material. For those parts, references will be provided to those sources.

1.2.1 What is Data Science?

According to Dr. Hector Bravo, Associate Professor of Computer Science at the University of Maryland, "Data science encapsulates the interdisciplinary activities required to create data-centric artifacts and applications that address specific scientific, socio-political, business, or other questions" (Bravo 2019). Dr. Bravo breaks down and defines this statement as follows: Data: Measurable units of information gathered or captured from activity of people, places and things (Bravo 2019). Specific Questions: Seeking to understand a phenomenon, natural, social or other, can we formulate specific questions for which an answer posed in terms of patterns observed, tested and or modeled in data is appropriate (Bravo 2019). Interdisciplinary activities: Formulating a question, assessing the appropriateness of the data and findings used to find an answer require understanding of the specific subject area. Deciding on the appropriateness of models and inferences made from models based on the data at hand requires understanding of statistical and computational methods (Bravo 2019). Data-Centric Artifacts and Applications: Answers to questions derived from data are usually shared and published in meaningful, succinct but sufficient, reproducible artifacts (papers, books, movies, comics). Going a step further, interactive applications that let others explore data, models and inferences are great (Bravo 2019).

Insights: More simply put, data science is getting data, organizing that data to be easily understood and processed, and making decisions from the data and its processing. Data science is important in most every field in the modern world. Examples include: Search engines like Google knowing exactly what we are searching for on the internet. Public health data being processed to analyze and prevent the outbreak of infectious diseases. Investigating what movies are the most popular in relation to box office gross. Data science can be used to analyze just about anything from any kind of topic.

1.2.2 General Workflow in Data Science

A simple organization of the data science workflow can be organized into the following parts:

Define the goal: What is the problem we are trying to solve?

Find and collect the data: What information needs to be collected? Where can that information be found?

Processing the data: From the data we have collected, what is actually needed? What is not needed and can be cleaned, or removed?

Exploring and analyzing the data: After cleaning the data, what patterns or trends can be found?

Show the results of the analysis: How can we easily represent the data and patterns that were found in the analysis? What models (ie. tables, graphs, or charts) could be used to show these results? Does the model(s) solve the problem?

Insights: As with any science, being successful relies on breaking down the necessary tasks into separate but related parts. Each of the parts in the data science process are interdependent, and therefore its deployment from beginning to end is imperative for success. A more in depth illustration of the Data Science Workflow can be found here -> http://www.hcbravo.org/IntroDataSci/bookdown-notes/introduction-and-overview.html#general-workflow (http://www.hcbravo.org/IntroDataSci/bookdown-notes/introduction-and-overview.html#general-workflow) From that link, a more complex illustration of the data science life cycle is copied here for reference (Bravo, 2019): http://www.hcbravo.org/IntroDataSci/bookdown-notes/img/zumel_mount_cycle.png (http://www.hcbravo.org/IntroDataSci/bookdown-notes/img/zumel_mount_cycle.png)

1.2.3 Transition to Crash Course

Sections 2-6 will illustrate the data science process using the World Happiness Report data -> https://www.kaggle.com/unsdsn/world-happiness (https://www.kaggle.com/unsdsn/world-happiness) set as an example.

2 Defining the Goal

In defining a goal, some key questions should be answered: What is the question or problem? What audience cares about this question or problem? How well can this problem or question be expected to be answered? How well does this audience expect this problem to be solved or question answered?

As an example, we've defined our own goal as follows: Happiness, or the pursuit there of, has been important to mankind since its early existence. So what factors contribute to people's happiness in the world? Is time a factor in people's happiness? Is wealth, government stability, or generosity a feature of happiness? Most people living in the world would care about answering this question. With the right data and analysis, this question can be generally answered for a population. We would think this audience (people in general) would have high expectations for this question to be answered.

Insights: Defining a goal and answering the above questions is the critical first step in starting the data science process. Without a defined goal, the remaining processes cannot take place.

3 Find and Collect the Data

A quick google search can yield an almost infinite amount of data sets on an infinite amount of topics and sites containing pertinent data sets. One such site, Kaggle, can be used to find data sets.

For our example on happiness in the world, Kaggle -> https://www.kaggle.com/ (https://www.kaggle.com/) was used to find a data set called World Happiness Report -> https://www.kaggle.com/unsdsn/world-happiness (https://www.kaggle.com/unsdsn/world-happiness) This data set is effective to answering our previously proposed

question about the factors influence happiness. Not only does this data set have variables for happiness rank and happiness score, but also for variables such as economy, family, health, freedom, government trust, and generosity, which could verify differences in happiness in a population.

```
temp_2015 <-
  read_csv("/Users/Ritesh/Documents/320/2015.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Region = col_character(),
##   HappinessRank = col_double(),
##   `Happiness Score` = col_double(),
##   `Standard Error` = col_double(),
##   `Economy (GDP per Capita)` = col_double(),
##   Family = col_double(),
##   `Health (Life Expectancy)` = col_double(),
##   Freedom = col_double(),
##   `Trust (Government Corruption)` = col_double(),
##   Generosity = col_double(),
##   `Dystopia Residual` = col_double()
## )
```

```
temp_2016 <-
  read_csv("/Users/Ritesh/Documents/320/2016.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Region = col_character(),
##   `Happiness Rank` = col_double(),
##   `Happiness Score` = col_double(),
##   `Lower Confidence Interval` = col_double(),
##   `Upper Confidence Interval` = col_double(),
##   `Economy (GDP per Capita)` = col_double(),
##   Family = col_double(),
##   `Health (Life Expectancy)` = col_double(),
##   Freedom = col_double(),
##   `Trust (Government Corruption)` = col_double(),
##   Generosity = col_double(),
##   `Dystopia Residual` = col_double()
## )
```

```
temp_2017 <-
  read_csv("/Users/Ritesh/Documents/320/2017.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Happiness.Rank = col_double(),
##   Happiness.Score = col_double(),
##   Whisker.high = col_double(),
##   Whisker.low = col_double(),
##   Economy..GDP.per.Capita. = col_double(),
##   Family = col_double(),
##   Health..Life.Expectancy. = col_double(),
##   Freedom = col_double(),
##   Generosity = col_double(),
##   Trust..Government.Corruption. = col_double(),
##   Dystopia.Residual = col_double()
## )
```

```
temp_2015
```

```
## # A tibble: 158 x 12
##    Country Region HappinessRank `Happiness Scor~ `Standard Error`
##    <chr>   <chr>          <dbl>            <dbl>            <dbl>
##  1 Switze~ Weste~             1             7.59           0.0341
##  2 Iceland Weste~             2             7.56           0.0488
##  3 Denmark Weste~             3             7.53           0.0333
##  4 Norway  Weste~             4             7.52           0.0388
##  5 Canada  North~             5             7.43           0.0355
##  6 Finland Weste~             6             7.41           0.0314
##  7 Nether~ Weste~             7             7.38           0.0280
##  8 Sweden  Weste~             8             7.36           0.0316
##  9 New Ze~ Austr~             9             7.29           0.0337
## 10 Austra~ Austr~            10             7.28           0.0408
## # ... with 148 more rows, and 7 more variables: `Economy (GDP per
## #   Capita)` <dbl>, Family <dbl>, `Health (Life Expectancy)` <dbl>,
## #   Freedom <dbl>, `Trust (Government Corruption)` <dbl>,
## #   Generosity <dbl>, `Dystopia Residual` <dbl>
```

```
temp_2016
```

```
## # A tibble: 157 x 13
##    Country Region `Happiness Rank` `Happiness Scor~ `Lower Confiden~
##    <chr>   <chr>             <dbl>            <dbl>            <dbl>
##  1 Denmark Weste~                1             7.53             7.46
##  2 Switze~ Weste~                2             7.51             7.43
##  3 Iceland Weste~                3             7.50             7.33
##  4 Norway  Weste~                4             7.50             7.42
##  5 Finland Weste~                5             7.41             7.35
##  6 Canada  North~                6             7.40             7.34
##  7 Nether~ Weste~                7             7.34             7.28
##  8 New Ze~ Austr~                8             7.33             7.26
##  9 Austra~ Austr~                9             7.31             7.24
## 10 Sweden  Weste~               10             7.29             7.23
## # ... with 147 more rows, and 8 more variables: `Upper Confidence
## #   Interval` <dbl>, `Economy (GDP per Capita)` <dbl>, Family <dbl>,
## #   `Health (Life Expectancy)` <dbl>, Freedom <dbl>, `Trust (Government
## #   Corruption)` <dbl>, Generosity <dbl>, `Dystopia Residual` <dbl>
```

```
temp_2017
```

```
## # A tibble: 155 x 12
##    Country Happiness.Rank Happiness.Score Whisker.high Whisker.low
##    <chr>            <dbl>           <dbl>        <dbl>       <dbl>
##  1 Norway               1            7.54         7.59        7.48
##  2 Denmark              2            7.52         7.58        7.46
##  3 Iceland              3            7.50         7.62        7.39
##  4 Switze~              4            7.49         7.56        7.43
##  5 Finland              5            7.47         7.53        7.41
##  6 Nether~              6            7.38         7.43        7.33
##  7 Canada               7            7.32         7.38        7.25
##  8 New Ze~              8            7.31         7.38        7.25
##  9 Sweden               9            7.28         7.34        7.22
## 10 Austra~             10            7.28         7.36        7.21
## # ... with 145 more rows, and 7 more variables:
## #   Economy..GDP.per.Capita. <dbl>, Family <dbl>,
## #   Health..Life.Expectancy. <dbl>, Freedom <dbl>, Generosity <dbl>,
## #   Trust..Government.Corruption. <dbl>, Dystopia.Residual <dbl>
```

The .csv files are downloaded locally, and RStudio is used to get and capture the data to be used in further processing and analysis. The read_csv command is used to read the World Happiness Report .csv file, and is subsequently assigned to table(s) or data.frame(s) depending on the year of the report (tab_2015, tab_2016, tab_2017). These table variables, if listed in the r code block, one per line, can be used to display the data collected from the .csv files.

A data.frame is a basic structure used to represent data (like a spreadsheet). Rows are observations (entities) while columns are variables that describe observations (attributes).

For more information on getting the data: Getting the Data -> http://www.hcbravo.org/IntroDataSci/bookdown-notes/measurements-and-data-types.html#getting-data (http://www.hcbravo.org/IntroDataSci/bookdown-notes/measurements-and-data-types.html#getting-data)

Insights: The internet is home to so much data that is ready and waiting to be used in further data science exploration and analysis. Any interest and research in minutes can give you an immense about of information that can be further collected into RStudio. RStudio is a powerful tool that can be used to further explore this data. Subsequent sections in this tutorial will show how to further process and analyze such data sets.

4 Processing the data

Before processing the collected data.frame(s), a few key RStudio operations and concepts involving one table can be found here for reference: Basic Operations -> http://www.hcbravo.org/IntroDataSci/bookdown-notes/principles-basic-operations.html (http://www.hcbravo.org/IntroDataSci/bookdown-notes/principles-basic-operations.html) and More Operations -> http://www.hcbravo.org/IntroDataSci/bookdown-notes/principles-more-operations.html (http://www.hcbravo.org/IntroDataSci/bookdown-notes/principles-more-operations.html).

```
tab_2015 <-
  select(temp_2015,-c(Region))
tab_2016 <-
  select(temp_2016,-c(Region))

tab_2015
```

```
## # A tibble: 158 x 11
##    Country HappinessRank `Happiness Scor~ `Standard Error` `Economy (GDP p~
##    <chr>           <dbl>            <dbl>            <dbl>            <dbl>
##  1 Switze~             1             7.59           0.0341             1.40
##  2 Iceland             2             7.56           0.0488             1.30
##  3 Denmark             3             7.53           0.0333             1.33
##  4 Norway              4             7.52           0.0388             1.46
##  5 Canada              5             7.43           0.0355             1.33
##  6 Finland             6             7.41           0.0314             1.29
##  7 Nether~             7             7.38           0.0280             1.33
##  8 Sweden              8             7.36           0.0316             1.33
##  9 New Ze~             9             7.29           0.0337             1.25
## 10 Austra~            10             7.28           0.0408             1.33
## # ... with 148 more rows, and 6 more variables: Family <dbl>, `Health
## #   (Life Expectancy)` <dbl>, Freedom <dbl>, `Trust (Government
## #   Corruption)` <dbl>, Generosity <dbl>, `Dystopia Residual` <dbl>
```

```
tab_2016
```

```
## # A tibble: 157 x 12
##    Country `Happiness Rank` `Happiness Scor~ `Lower Confiden~
##    <chr>              <dbl>            <dbl>            <dbl>
##  1 Denmark                1             7.53             7.46
##  2 Switze~                2             7.51             7.43
##  3 Iceland                3             7.50             7.33
##  4 Norway                 4             7.50             7.42
##  5 Finland                5             7.41             7.35
##  6 Canada                 6             7.40             7.34
##  7 Nether~                7             7.34             7.28
##  8 New Ze~                8             7.33             7.26
##  9 Austra~                9             7.31             7.24
## 10 Sweden                10             7.29             7.23
## # ... with 147 more rows, and 8 more variables: `Upper Confidence
## #   Interval` <dbl>, `Economy (GDP per Capita)` <dbl>, Family <dbl>,
## #   `Health (Life Expectancy)` <dbl>, Freedom <dbl>, `Trust (Government
## #   Corruption)` <dbl>, Generosity <dbl>, `Dystopia Residual` <dbl>
```

In this block of code, we filtered the data, removing the region attribute as it was not pertinent to the data analysis (we only needed the country not the region).

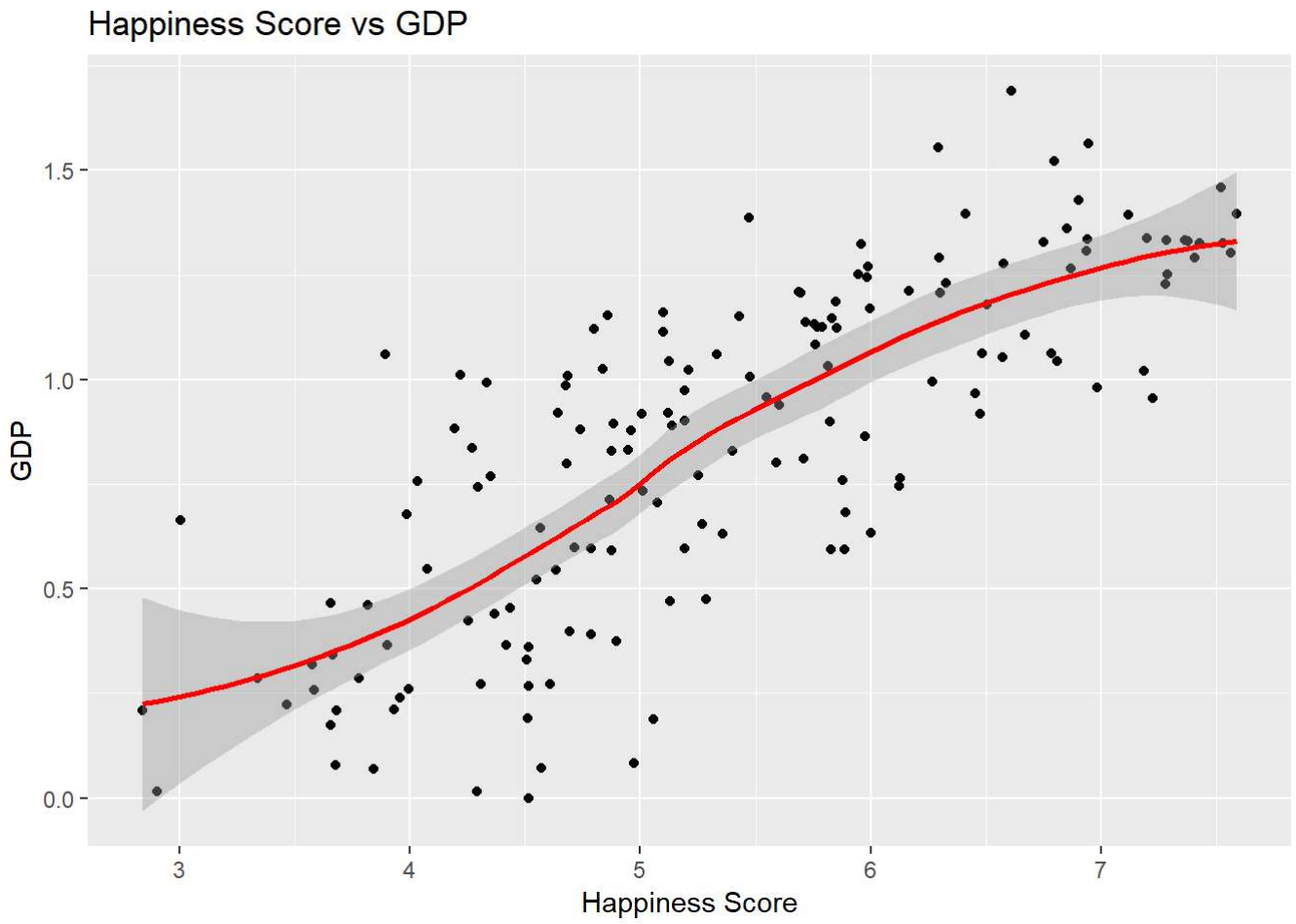
Insights: Most data sets can have millions (if not billions and trillions) of different pieces data, of which is not pertinent to the current goal. It is therefore advantageous to process the data, removing information that will only clutter, confuse, and reduce the efficiency of processing the data.

5 Exploring and analyzing the data

The goal of Exploratory Data Analysis (EDA) is to perform an initial exploration of attributes/variables across entities/observations. Please reference this -> http://www.hcbravo.org/IntroDataSci/bookdown-notes/exploratory-data-analysis-visualization.html (http://www.hcbravo.org/IntroDataSci/bookdown-notes/exploratory-data-analysis-visualization.html) for an explanation on this process.

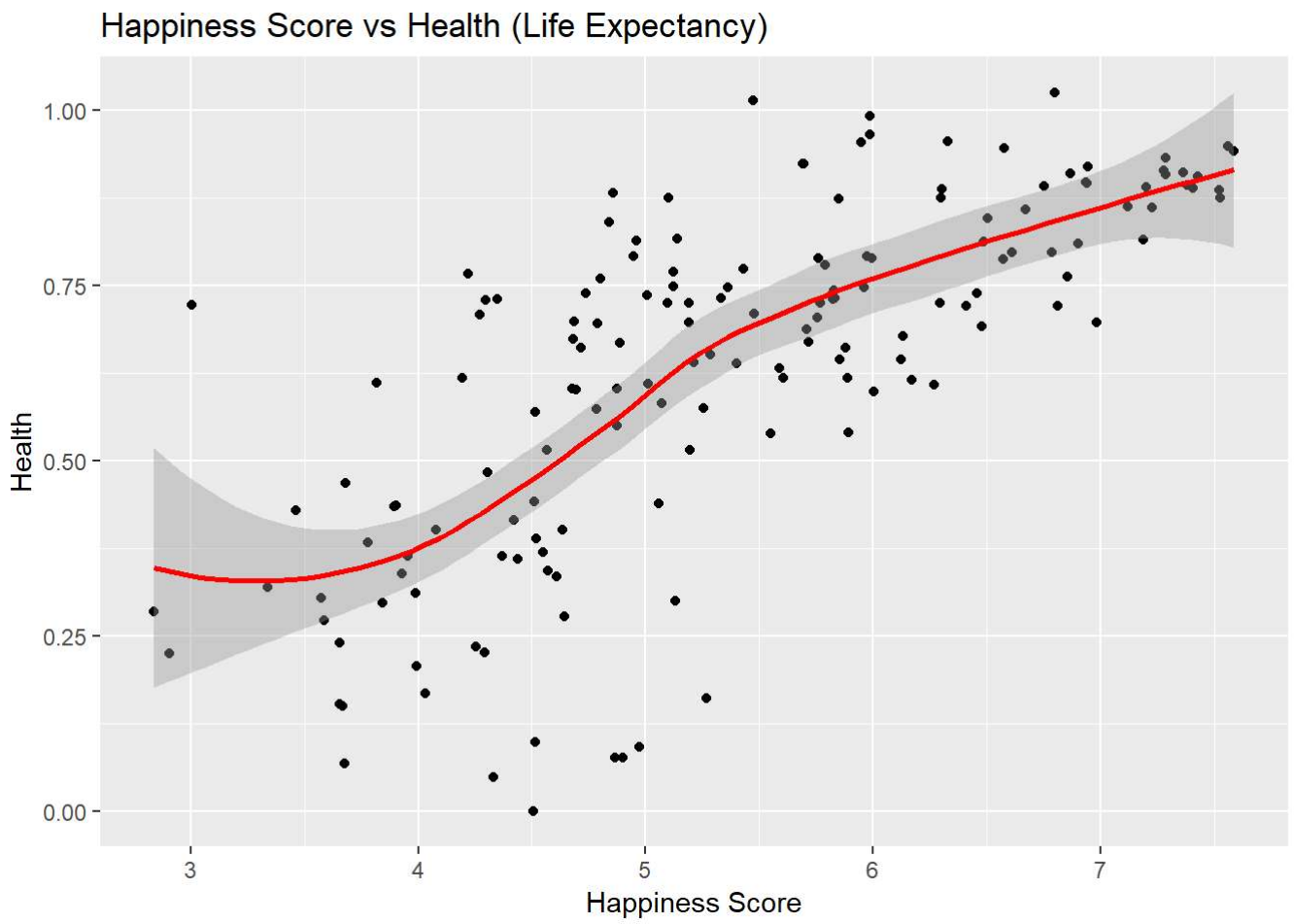
```
scoreVGDP <- tab_2015 %>%
  ggplot(aes(x = `Happiness Score`, y = `Economy (GDP per Capita)`)) +
  geom_point() +
  geom_smooth(method='auto', color = "red") +
  ggtitle("Happiness Score vs GDP") +
  labs(y="GDP") +
  labs(x="Happiness Score")
scoreVGDP
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



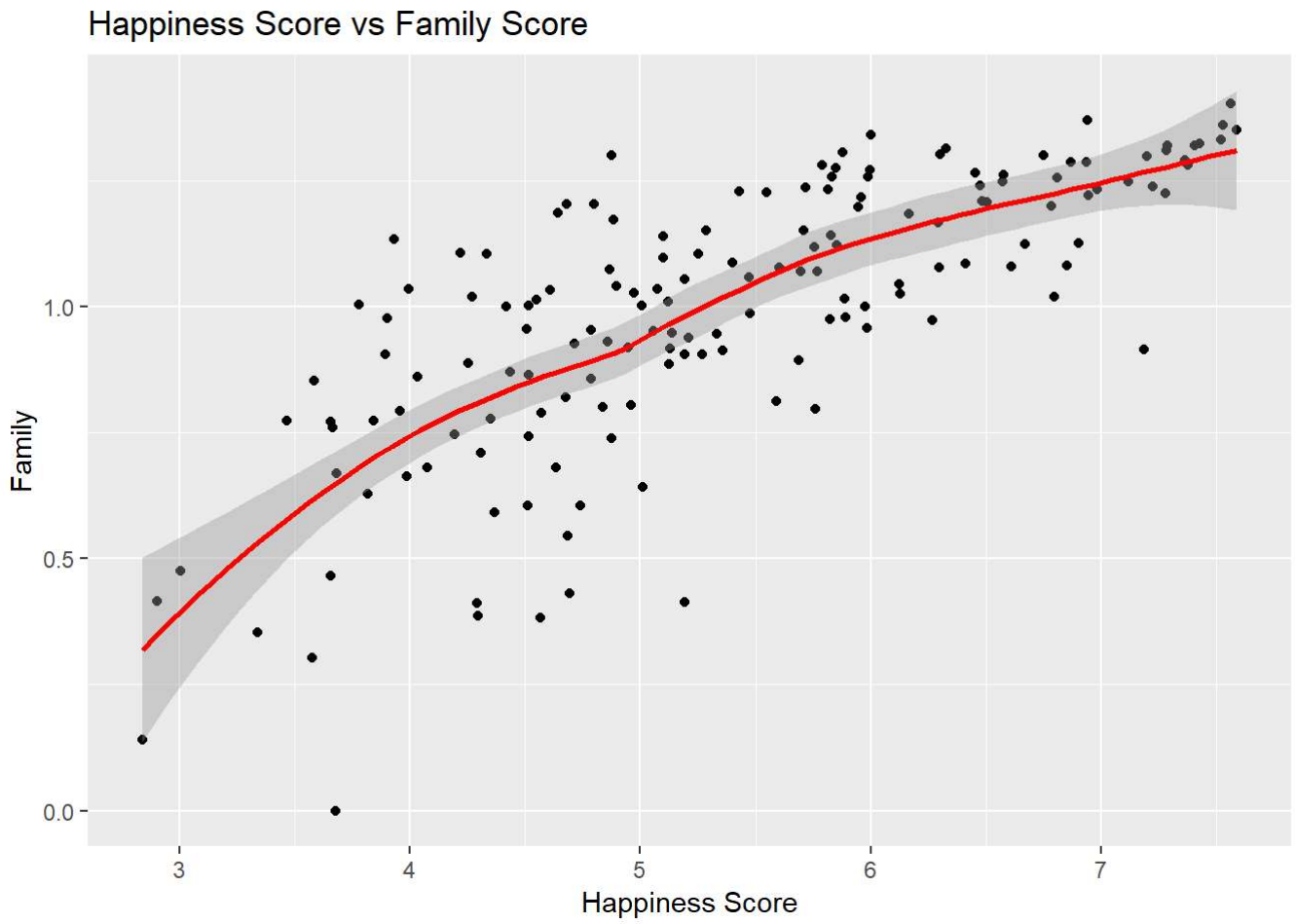

```
scoreVHLT <- tab_2015 %>%
  ggplot(aes(x=`Happiness Score`, y=`Health (Life Expectancy)`)) +
  geom_point() +
  geom_smooth(method='auto', color = "red") +
  ggtitle("Happiness Score vs Health (Life Expectancy)") +
  labs(y="Health") +
  labs(x="Happiness Score")
scoreVHLT
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



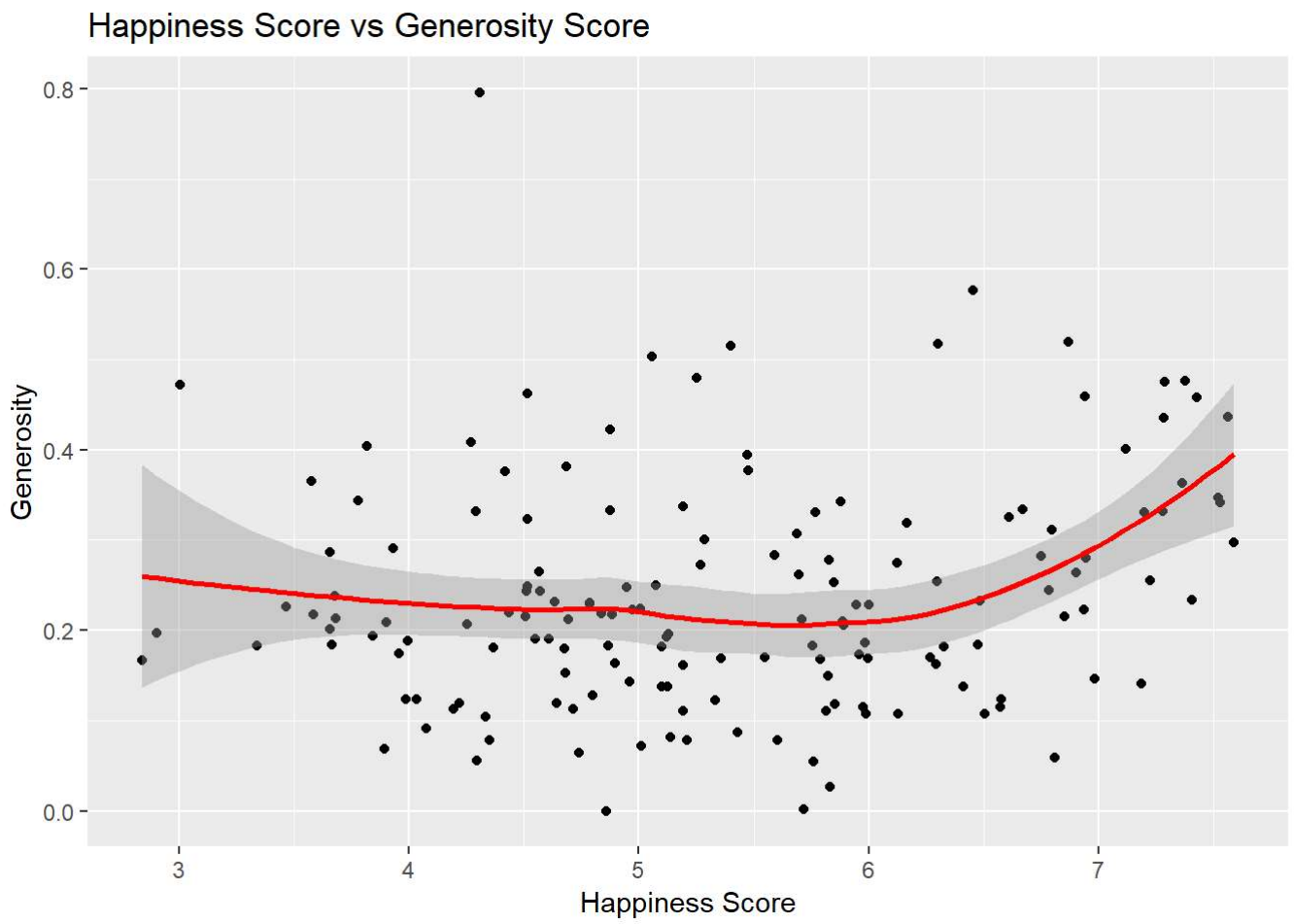

```
scoreVFm <- tab_2015 %>%
  ggplot(aes(x = `Happiness Score`, y = Family)) +
  geom_point() +
  geom_smooth(method='auto', color = "red") +
  ggtitle("Happiness Score vs Family Score") +
  labs(y="Family") +
  labs(x="Happiness Score")
scoreVFm
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



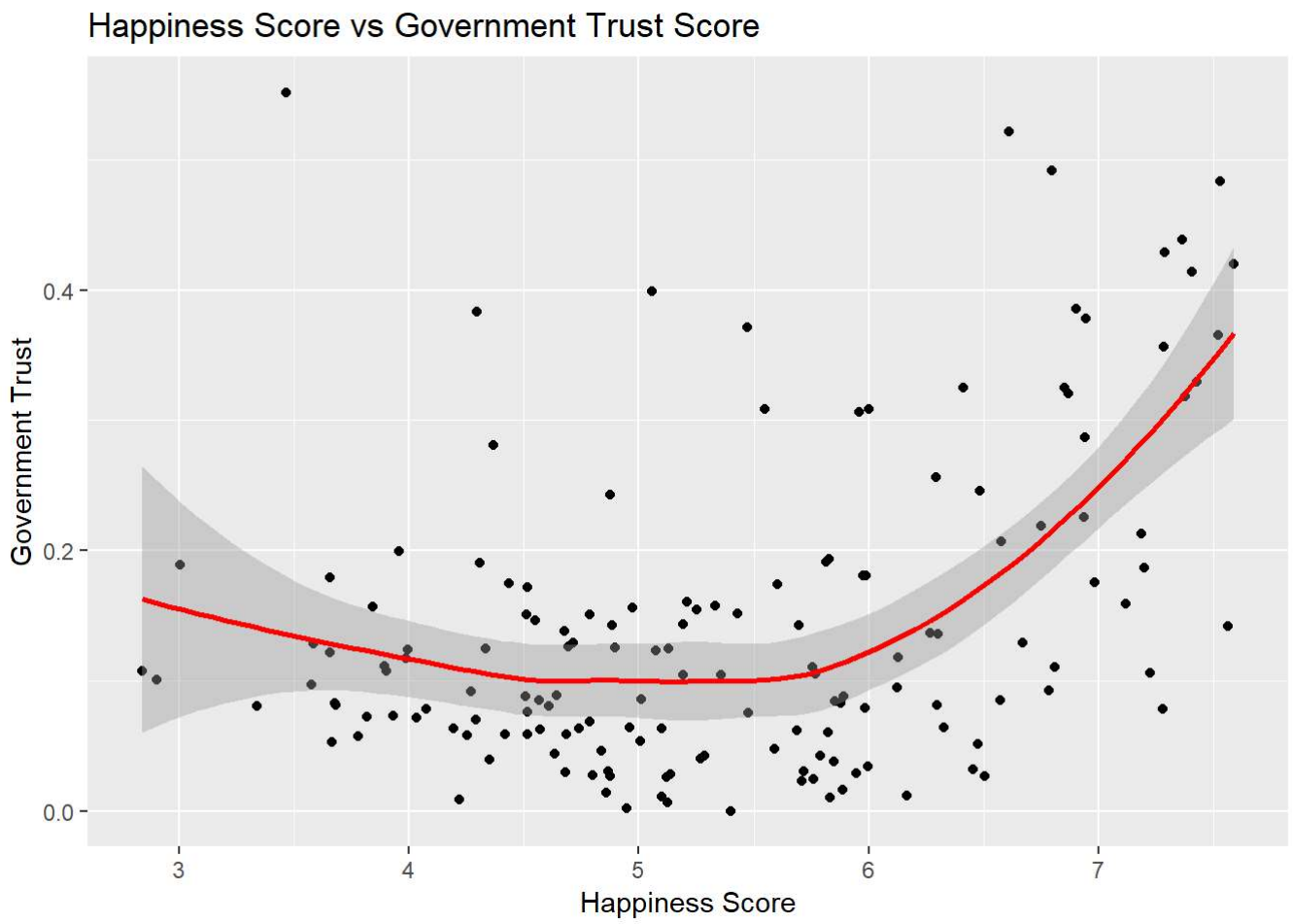

```
scoreVGen <- tab_2015 %>%
  ggplot(aes(x = `Happiness Score`, y = Generosity)) +
  geom_point() +
  geom_smooth(method='auto', color = "red") +
  ggtitle("Happiness Score vs Generosity Score") +
  labs(y="Generosity") +
  labs(x="Happiness Score")
scoreVGen
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



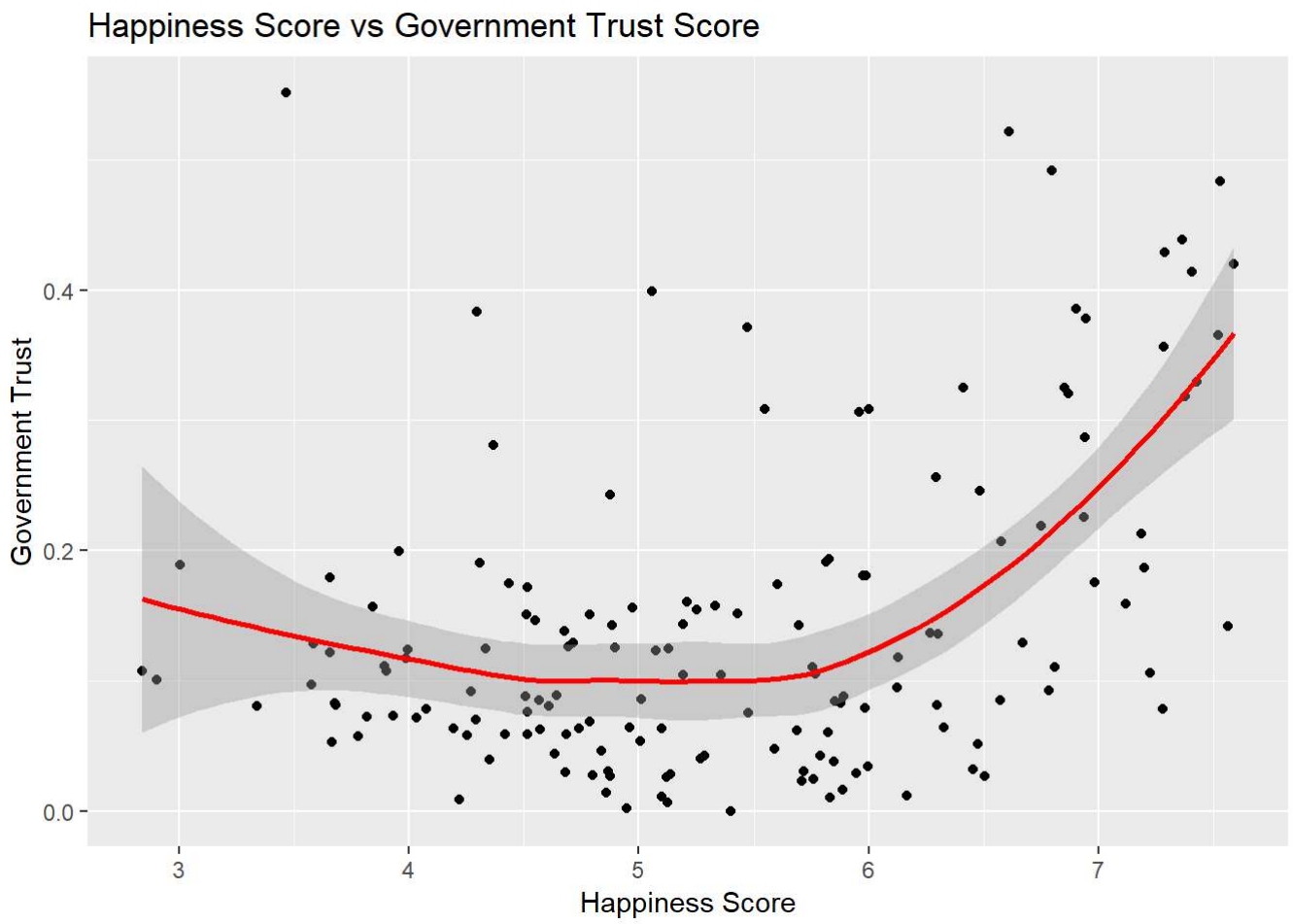

```
scoreVTrust <- tab_2015 %>%
  ggplot(aes(x = `Happiness Score`, y = `Trust (Government Corruption)`)) +
  geom_point() +
  geom_smooth(method='auto', color = "red") +
  ggtitle("Happiness Score vs Government Trust Score") +
  labs(y="Government Trust") +
  labs(x="Happiness Score")
scoreVTrust
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



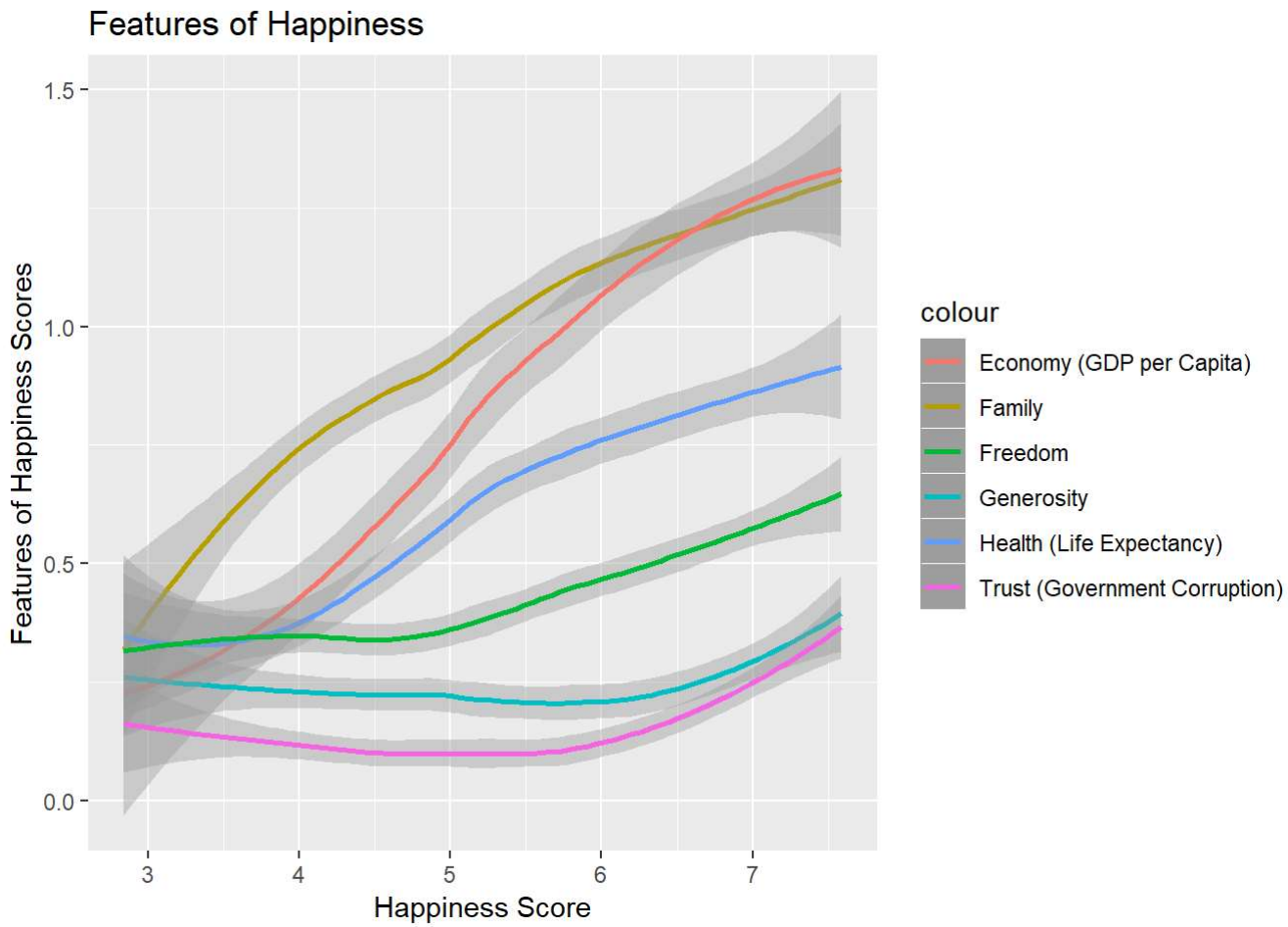

```
scoreVFreedom <- tab_2015 %>%
  ggplot(aes(x = `Happiness Score`, y = Freedom)) +
  geom_point() +
  geom_smooth(method='auto', color = "red") +
  ggtitle("Happiness Score vs Freedom Score") +
  labs(y="Freedom Trust") +
  labs(x="Happiness Score")
scoreVTrust
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
scoreAll <- tab_2015 %>%
  ggplot(aes(x=`Happiness Score`)) +
  geom_smooth(method='auto',aes(y=Family, color="Family")) +
  geom_smooth(method='auto',aes(y=`Economy (GDP per Capita)` ,color="Economy (GDP per Capita)"))
+
  geom_smooth(method='auto',aes(y=Generosity,  color="Generosity")) +
  geom_smooth(method='auto',aes(y=`Health (Life Expectancy)`,  color="Health (Life Expectancy)"
)) +
  geom_smooth(method='auto',aes(y=`Trust (Government Corruption)`,  color="Trust (Government Cor
ruption)")) +
  geom_smooth(method='auto',aes(y=Freedom, color="Freedom")) +
  ggtitle("Features of Happiness") +
  labs(y="Features of Happiness Scores") +
  labs(x="Happiness Score")
scoreAll
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

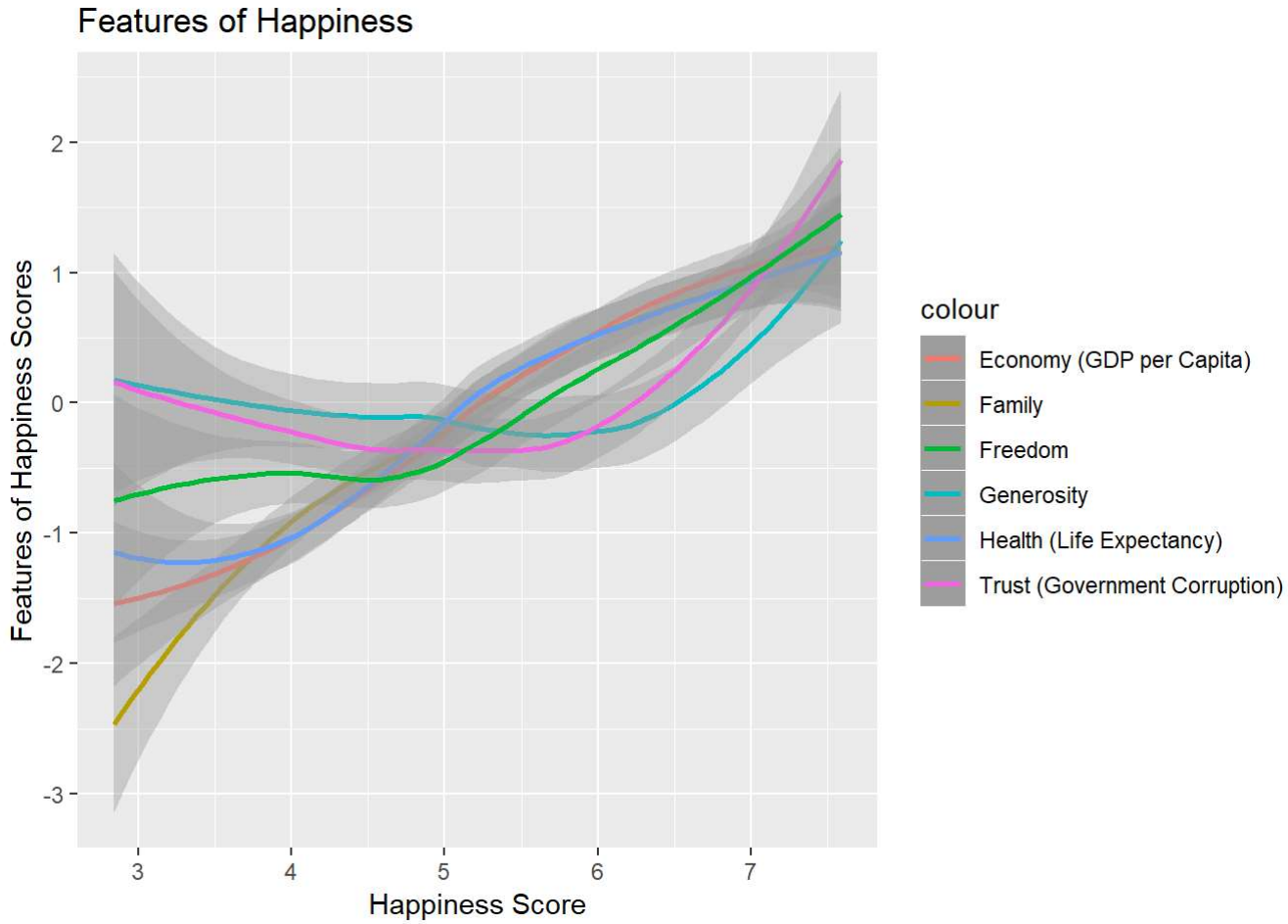


For this section of analysis, we first plotted data for each feature of happiness (y value) by happiness score (x value) for the 2015 year. Separately, this data does not show much, so we combined all happiness features into one plot. We then realized that we needed to normalize the data, which is as follows:

```
tab_2015$`Health (Life Expectancy)` <- scale(tab_2015$`Health (Life Expectancy)`)
tab_2015$`Economy (GDP per Capita)` <- scale(tab_2015$`Economy (GDP per Capita)`)
tab_2015$Family <- scale(tab_2015$Family)
tab_2015$Generosity <- scale(tab_2015$Generosity)
tab_2015$`Trust (Government Corruption)` <- scale(tab_2015$`Trust (Government Corruption)`)
tab_2015$`Freedom` <- scale(tab_2015$`Freedom`)

scoreAll <- tab_2015 %>%
ggplot(aes(x=`Happiness Score`)) +
geom_smooth(method='auto',aes(y=Family, color="Family")) +
geom_smooth(method='auto',aes(y=`Economy (GDP per Capita)` , color="Economy (GDP per Capita)"))
 +
geom_smooth(method='auto',aes(y=Generosity, color="Generosity")) +
geom_smooth(method='auto',aes(y=`Health (Life Expectancy)`, color="Health (Life Expectancy)")) +
geom_smooth(method='auto',aes(y=`Trust (Government Corruption)`, color="Trust (Government Corrup
tion)")) +
geom_smooth(method='auto',aes(y=Freedom, color="Freedom")) +
ggtitle("Features of Happiness") +
labs(y="Features of Happiness Scores") +
labs(x="Happiness Score")
scoreAll
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



In order to properly vizualize the data, normalizing the data was key so that the units of each feature do not interact with our interpretation of their influence on happiness. That is where we found the normalized part in our conclusions.

Conclusions of our data shows that we can see that in the context of happiness, economy, family, and health are strongly coorelated with each other, while generosity and trust are strongly correlated with each other. Freedom on the otherhand seems to be related to both sets of groups. Overall, the happier a country is the more prevalent each of the features they will have.

To answer our initial question, there are certain features that show an association between happiness.

Insights: Exploring and analyzing the data is the stage of data science that allows the data you have collected and cleaned to be transformed and manipulated into something that can be further understand and analyzed past the information collected. The information in and of itself is important, but turning it into something more informative is key to the data science process.

6 Further Visualization and Publishing of Results in RStudio

6.1 Visualizing Results in RStudio

This section describes a mean to further produce the data in various forms to express both the data and the results of the data analysis. As described previously, displaying the data frame as a table is one such method in visualizing the data. A more robust and practical way to visualize the data is in the form of a chart, graph, or plot. References to building charts, graphs, and plots in RStudio can be found here -> http://www.hcbravo.org/IntroDataSci/bookdown-notes/basic-plotting-with-ggplot.html (http://www.hcbravo.org/IntroDataSci/bookdown-notes/basic-plotting-with-ggplot.html) using the ggplot package in RStudio. We will be using this package to visualize our findings with the World Happiness Report data set.

```
top_2015 <-
  tab_2015 %>%
  filter(`HappinessRank` <= 10)
top_2016 <-
  tab_2016 %>%
  filter(`Happiness Rank` <= 10)
top_2017 <-
  temp_2017 %>%
  filter(Happiness.Rank <= 10)

top_2015
```

```
## # A tibble: 10 x 11
##    Country HappinessRank `Happiness Scor~ `Standard Error` `Economy (GDP p~
##    <chr>           <dbl>            <dbl>             <dbl>            <dbl>
##  1 Switze~             1             7.59           0.0341             1.37
##  2 Iceland             2             7.56           0.0488             1.13
##  3 Denmark             3             7.53           0.0333             1.19
##  4 Norway              4             7.52           0.0388             1.52
##  5 Canada              5             7.43           0.0355             1.19
##  6 Finland             6             7.41           0.0314             1.10
##  7 Nether~             7             7.38           0.0280             1.20
##  8 Sweden              8             7.36           0.0316             1.20
##  9 New Ze~             9             7.29           0.0337             1.00
## 10 Austra~            10             7.28           0.0408             1.21
## # ... with 6 more variables: Family[,1] <dbl>, `Health (Life
## #   Expectancy)`[,1] <dbl>, Freedom[,1] <dbl>, `Trust (Government
## #   Corruption)`[,1] <dbl>, Generosity[,1] <dbl>, `Dystopia
## #   Residual` <dbl>
```

```
top_2016
```

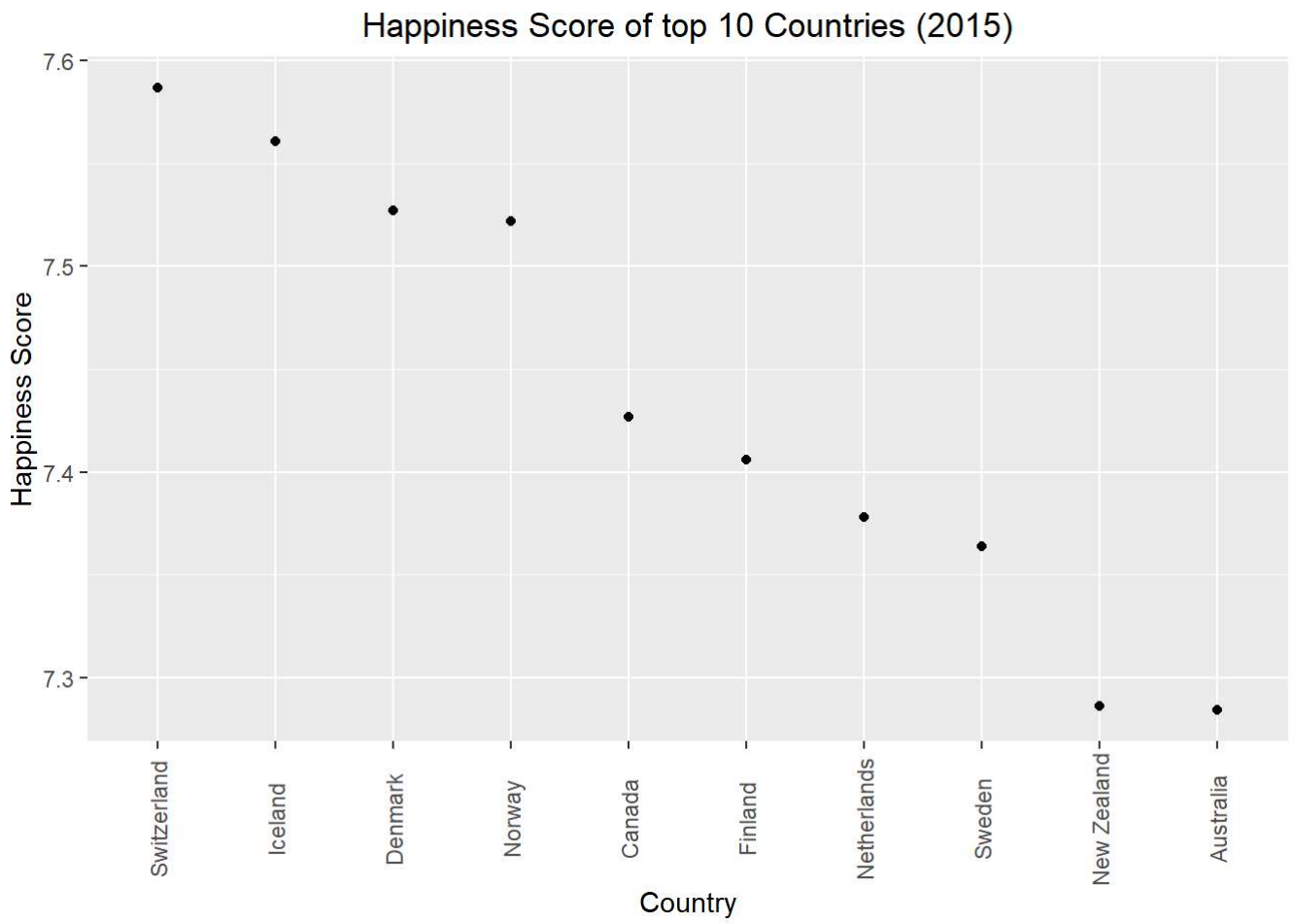
```
## # A tibble: 10 x 12
##    Country `Happiness Rank` `Happiness Scor~ `Lower Confiden~
##    <chr>              <dbl>            <dbl>            <dbl>
##  1 Denmark                1             7.53             7.46
##  2 Switze~                2             7.51             7.43
##  3 Iceland                3             7.50             7.33
##  4 Norway                 4             7.50             7.42
##  5 Finland                5             7.41             7.35
##  6 Canada                 6             7.40             7.34
##  7 Nether~                7             7.34             7.28
##  8 New Ze~                8             7.33             7.26
##  9 Austra~                9             7.31             7.24
## 10 Sweden                10             7.29             7.23
## # ... with 8 more variables: `Upper Confidence Interval` <dbl>, `Economy
## #   (GDP per Capita)` <dbl>, Family <dbl>, `Health (Life
## #   Expectancy)` <dbl>, Freedom <dbl>, `Trust (Government
## #   Corruption)` <dbl>, Generosity <dbl>, `Dystopia Residual` <dbl>
```

```
top_2017
```

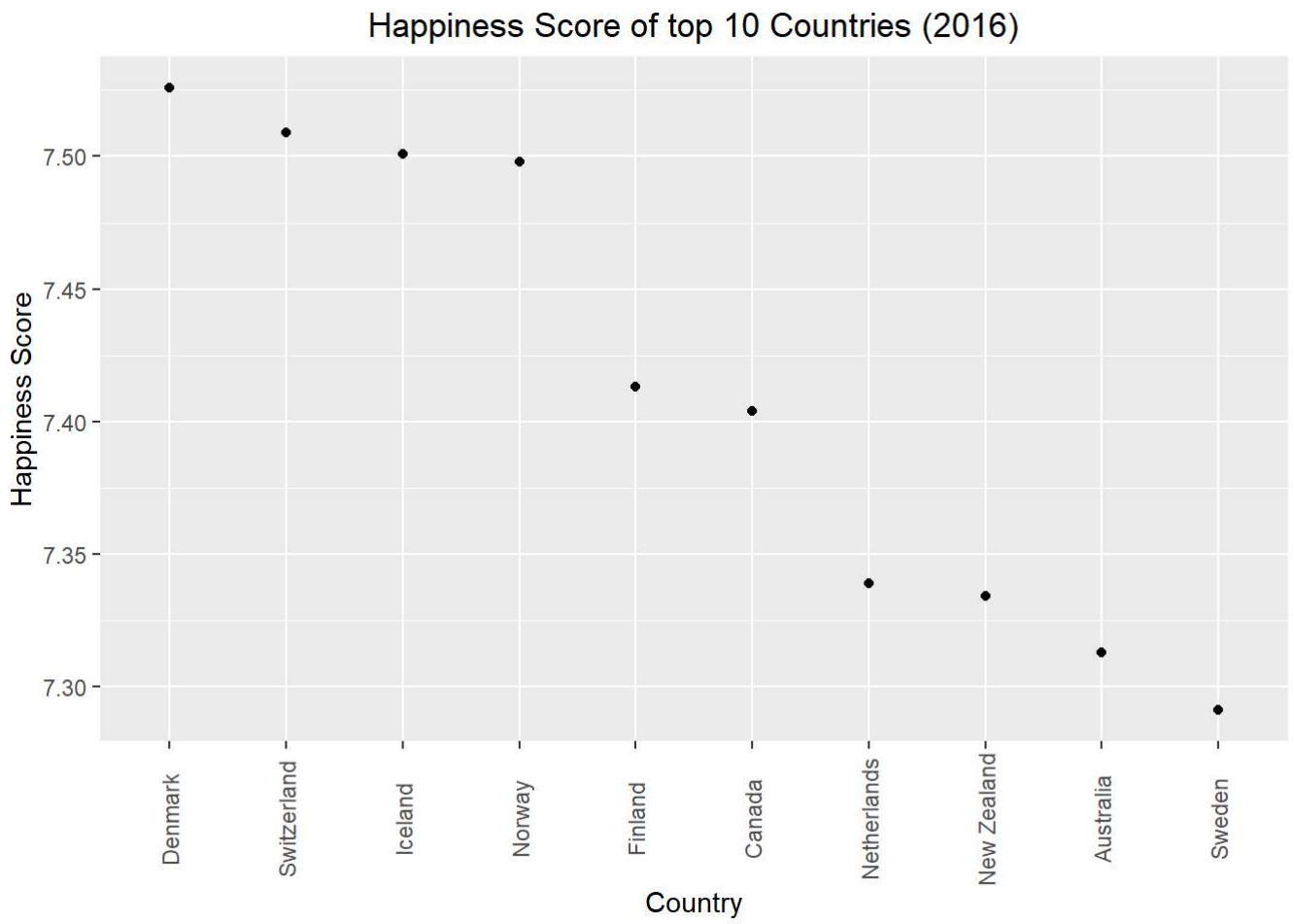
```
## # A tibble: 10 x 12
##    Country Happiness.Rank Happiness.Score Whisker.high Whisker.low
##    <chr>            <dbl>           <dbl>        <dbl>       <dbl>
##  1 Norway               1            7.54         7.59        7.48
##  2 Denmark              2            7.52         7.58        7.46
##  3 Iceland              3            7.50         7.62        7.39
##  4 Switze~              4            7.49         7.56        7.43
##  5 Finland              5            7.47         7.53        7.41
##  6 Nether~              6            7.38         7.43        7.33
##  7 Canada               7            7.32         7.38        7.25
##  8 New Ze~              8            7.31         7.38        7.25
##  9 Sweden               9            7.28         7.34        7.22
## 10 Austra~             10            7.28         7.36        7.21
## # ... with 7 more variables: Economy..GDP.per.Capita. <dbl>, Family <dbl>,
## #   Health..Life.Expectancy. <dbl>, Freedom <dbl>, Generosity <dbl>,
## #   Trust..Government.Corruption. <dbl>, Dystopia.Residual <dbl>
```

Using additional r code, we have created 3 new tables holding just the top 10 countries by happiness rating.

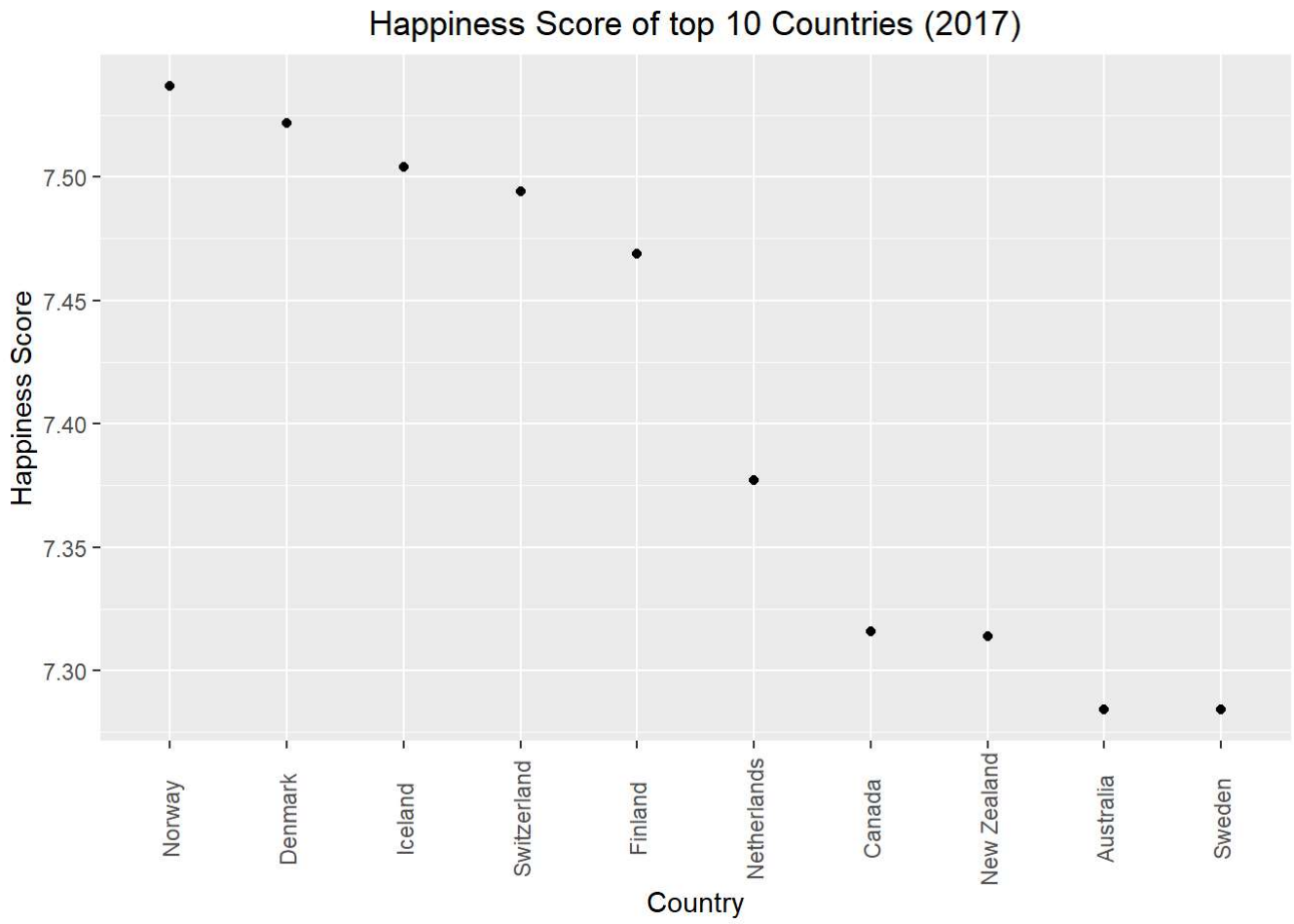
```
top_2015 %>%
  ggplot(aes(x=reorder(Country, -`Happiness Score`), y=`Happiness Score`)) +
  geom_point() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5), plot.title = element_text(hjust = 0.5))
+
  ggtitle("Happiness Score of top 10 Countries (2015)") +
  labs(x="Country") +
  labs(y="Happiness Score")
```




```
top_2016 %>%
  ggplot(aes(x=reorder(Country, -`Happiness Score`), y=`Happiness Score`)) +
  geom_point() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5), plot.title = element_text(hjust = 0.5))
+
  ggtitle("Happiness Score of top 10 Countries (2016)") +
  labs(x="Country") +
  labs(y="Happiness Score")
```




```
top_2017 %>%
  ggplot(aes(x=reorder(Country, -Happiness.Score), y=Happiness.Score)) +
  geom_point() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5), plot.title = element_text(hjust = 0.5))
+
  ggtitle("Happiness Score of top 10 Countries (2017)") +
  labs(x="Country") +
  labs(y="Happiness Score")
```



The three plots show the 10 highest countries and their happiness scores, which simply put, have a decreasing trend of happiness scores. Three plots are made for 2015, 2016, and 2017.

6.2 Publishing Results Using RStudio

An impressive and effective publishing tool known as Rmarkdown can be used in RStudio to publish the work and findings made in an RStudio project. A brief introduction to using Rmarkdown can be found here -> http://www.hcbravo.org/IntroDataSci/bookdown-notes/brief-introduction-to-rmarkdown.html (http://www.hcbravo.org/IntroDataSci/bookdown-notes/brief-introduction-to-rmarkdown.html)

Insights: Properly visualizing and publishing your data science projects is almost as critical as coming up with the initial goal because if no one can properly see your results than they will not have much to say or learn about it. Therefore, it may not even exist. That is why the operations and procedures in this section are of critical importance as conveying the data you have analyzed is critical to bringing your results to your audience.

7 Tutorial Conclusion

I hope you found this simple crash course tutorial to data science not only informative, but interesting! The world of data science is vast and is ready for you to explore and analyze it!