The Art of Storytelling in Analytics and Data Science | How to Create Data Stories?

BEGINNER BUSINESS ANALYTICS DATA VISUALIZATION R RESOURCE STRUCTURED DATA

Introduction

The idea of storytelling is fascinating; to take an idea or an incident, and turn it into a story. It brings the idea to life and makes it more interesting. This happens in our day to day life. Whether we narrate a funny incident or our findings, stories have always been the “go-to” to draw interest from listeners and readers alike.

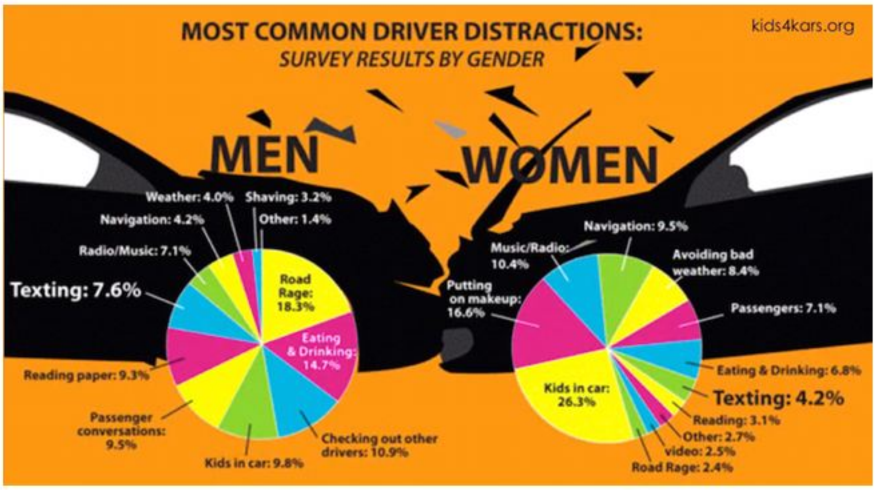
For instance; when we talk of how one of our friends got scolded by a teacher, we tend to narrate the incident from the beginning so that a flow is maintained.

Let’s take an example of the most common driving distractions by gender. There are two ways to tell this.

The first is that I give you some statistics as follows:

1. **6% of men** believe texting is a distraction as compared to **4.2% of the women**.
2. Kids in the car cause **9.8% of the men** to be distracted as compared to **26.3% of the women**.

Another way to recreate similar statistics is this visual from kids4kars.org.



Which one do you think tells a better story?

Note: Make sure you check out the comprehensive multi-course Certified Business Analytics Program that covers the art of storytelling through various industry examples and using tools like Excel, Python and Tableau.

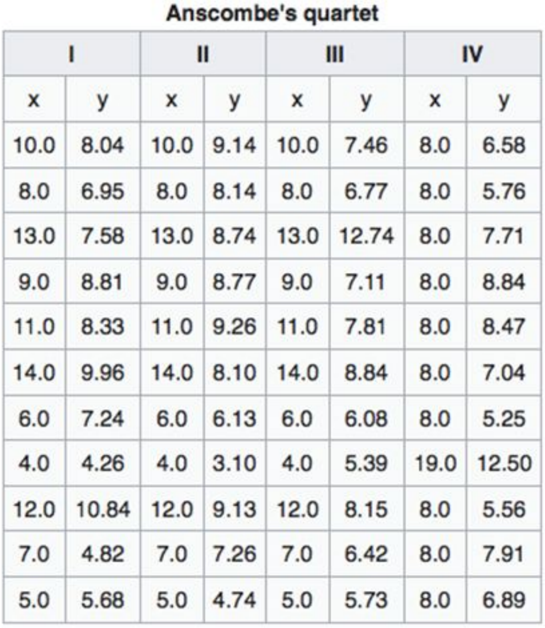
Table of Contents

The Need for storytelling

The art of storytelling is simple and complex at the same time. Stories provoke thought and bring out insights that could not have been understood or explained before. It's often overlooked in data-driven operations as we believe it's a trivial task.
What we fail to understand is that the best stories not presented well end up being useless!

In several firms, the first step towards analyzing anything is story-boarding. Questions like why do we have to analyze it? what decisions can we make out of it? Sometimes, data alone tells such visual and intricate stories that we don't need to run complex correlations to confirm it.

The best example of needing stories and visuals to explain data is the Anscombe's Quartet. The Anscombe's Quartet is a set of four datasets with very similar statistical summaries, but completely different when you visualize them.

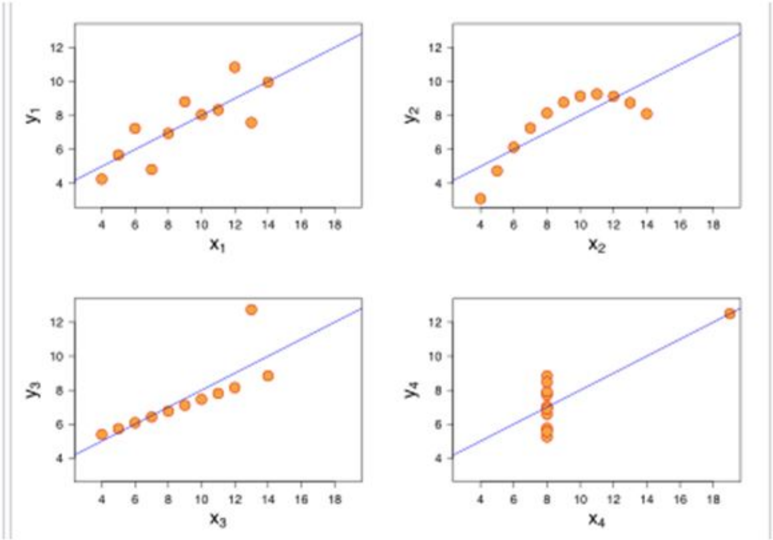


Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

These are the four datasets used during the depiction of the Anscombe's Quartet. If we look at mere numbers, we find that their summary statistics are almost identical.

Let's see how they appear when we visualize them.

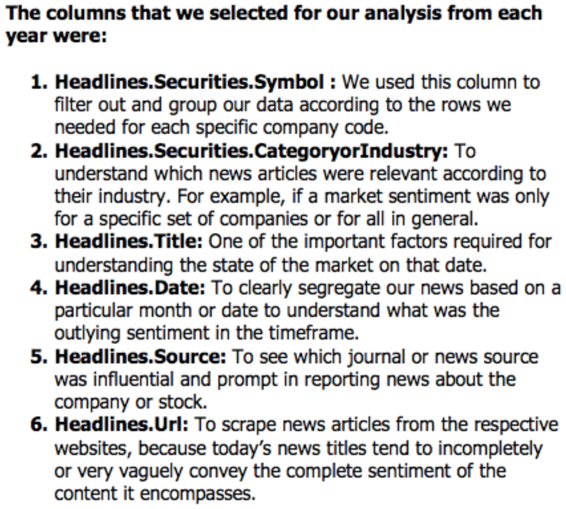


Did you ever think these four quartets would have such varying visuals?

How to create stories?

To create a story or a plot is the first step to selling your ideas with a strong foot forward. Most people fail to think their stories through and cannot differentiate themselves from mediocrity. Let me take an example and guide you through the steps of creating stories.

We will be exploring a dataset that has news headlines and details of every stock price from the NASDAQ 100 tech companies. The columns selected are as follows.



The columns that we selected for our analysis from each year were:

1. **Headlines.Securities.Symbol :** We used this column to filter out and group our data according to the rows we needed for each specific company code.
2. **Headlines.Securities.CategoryorIndustry:** To understand which news articles were relevant according to their industry. For example, if a market sentiment was only for a specific set of companies or for all in general.
3. **Headlines.Title:** One of the important factors required for understanding the state of the market on that date.
4. **Headlines.Date:** To clearly segregate our news based on a particular month or date to understand what was the outlying sentiment in the timeframe.
5. **Headlines.Source:** To see which journal or news source was influential and prompt in reporting news about the company or stock.
6. **Headlines.Url:** To scrape news articles from the respective websites, because today's news titles tend to incompletely or very vaguely convey the complete sentiment of the content it encompasses.

1. Begin with a pen-paper approach

Visually engaging presentations will inspire your audience, but they definitely need more work to be put in. One of the best presentations have been created on rough pages and tissue papers.

Scripting down your ideas and flow before you start structuring your story is very essential to your final product.

The single most important thing you can do to dramatically improve your analytics is to have a story to tell. A flow that you can generate can have a lot of friction in your end result.

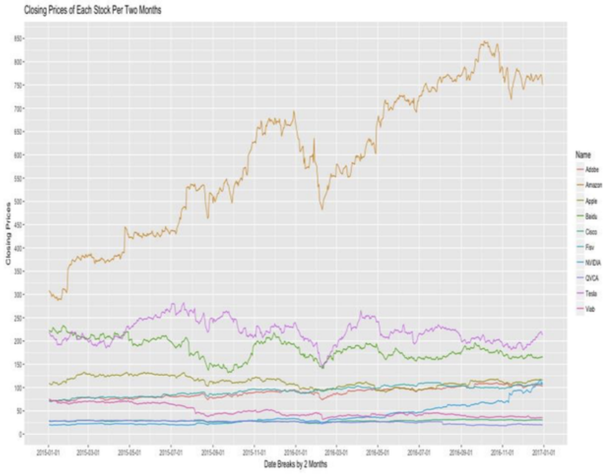
Aristotle's classic five-point plan that helps deliver strong impacts is:

1. Deliver a story or statement that arouses the audience's interest.
2. Pose a problem or question that has to be solved or answered.
3. Offer a solution to the problem you raised.
4. Describe specific benefits for adopting the course of action set forth in your solution.
5. State a call to action.

The way I structured my report was by involving plots that would give me a better understanding of my data.

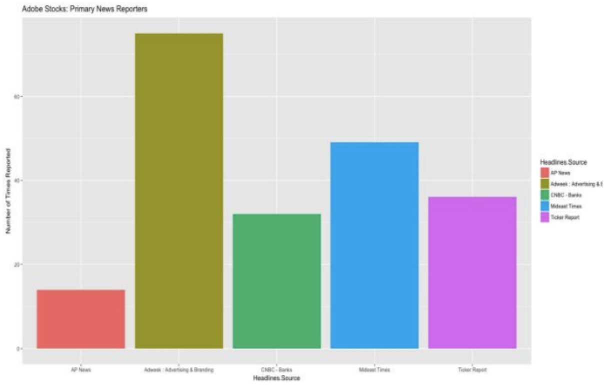
The first idea that I had was, how can I make better business decisions of stocks by using the data that I have?

1. Line Graph of each stock in a date period of two months.



Involving a line graph would help me analyze trend lines of specific stock prices.

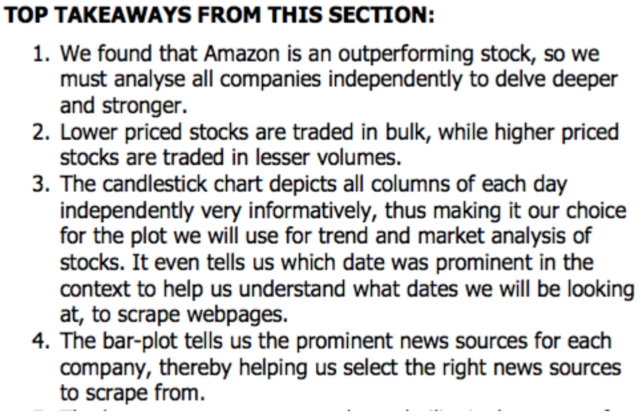
As I can see, February 2016 has been a drop for all stocks. This would help me scrape news articles only from that period to identify what caused the drop. Now, how do I select which news source to scrape from?



By identifying which news source reported most about a particular stock, we would have reason to believe that this is a good source for the specific stock.

2. Dig deeper to identify the sole purpose of your story

- Identify closely, what the idea of your story is. Ask yourself, “What am I really giving with this story?” It’s never the story alone, but what the story can do to make decision making better. What you’re displaying is the idea of a better decision making or analytics.
- Develop a personal “passion statement.” In one sentence, tell your prospects and why you are genuinely excited about working with them. Your passion statement will be remembered long.



TOP TAKEAWAYS FROM THIS SECTION:

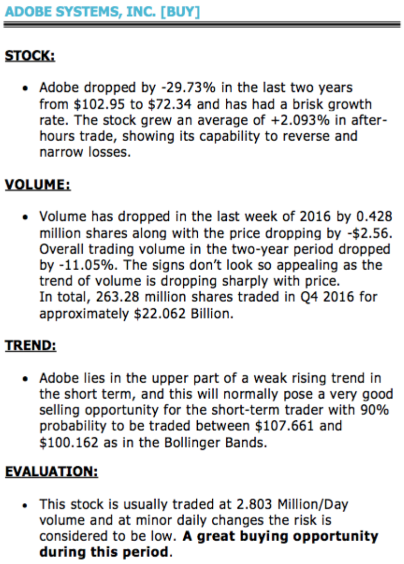
1. We found that Amazon is an outperforming stock, so we must analyse all companies independently to delve deeper and stronger.
2. Lower priced stocks are traded in bulk, while higher priced stocks are traded in lesser volumes.
3. The candlestick chart depicts all columns of each day independently very informatively, thus making it our choice for the plot we will use for trend and market analysis of stocks. It even tells us which date was prominent in the context to help us understand what dates we will be looking at, to scrape webpages.
4. The bar-plot tells us the prominent news sources for each company, thereby helping us select the right news sources to scrape from.

3. Use powerful headings

- Create your heading, a one-sentence statement for your story, visual, or analysis. The most effective headlines are concise, specific, and offer a personal benefit.
- Remember, your heading is a statement that offers your audience a vision of a better understanding. It’s not about you. It’s about them.

4. Design a Road-Map

- Create a list of all the key points you want your audience to know about your story, visual, or analysis.
- Categorize the list until you are left with only three major message points. This group of three will provide the verbal road map for your story.
- Under each of your three key messages, add supporting evidence to enhance the narrative. These could include some or all of the following: personal stories, facts, examples, analogies etc.



ADOBE SYSTEMS, INC. [BUY]

STOCK:

- Adobe dropped by -29.73% in the last two years from $102.95 to $72.34 and has had a brisk growth rate. The stock grew an average of +2.093% in after-hours trade, showing its capability to reverse and narrow losses.

VOLUME:

- Volume has dropped in the last week of 2016 by 0.428 million shares along with the price dropping by -$2.56. Overall trading volume in the two-year period dropped by -11.05%. The signs don't look so appealing as the trend of volume is dropping sharply with price.
  In total, 263.28 million shares traded in Q4 2016 for approximately $22.062 Billion.

TREND:

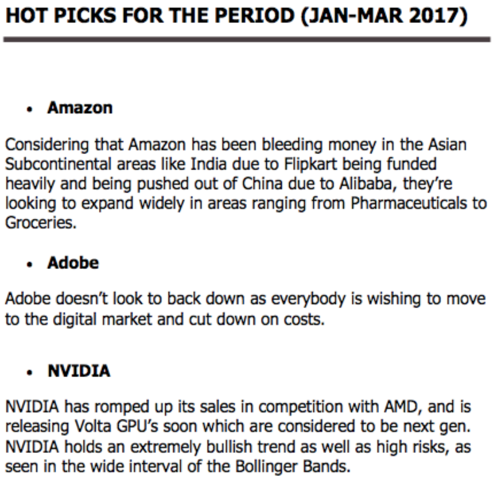
- Adobe lies in the upper part of a weak rising trend in the short term, and this will normally pose a very good selling opportunity for the short-term trader with 90% probability to be traded between $107.661 and $100.162 as in the Bollinger Bands.

EVALUATION:

- This stock is usually traded at 2.803 Million/Day volume and at minor daily changes the risk is considered to be low. **A great buying opportunity during this period**.

5. Conclude with brevity

Now that you have put forward all points of your story, your conclusion should be short and powerful. In my report, I mentioned small 3-4 liner summaries to conclude why to buy a particular stock.



HOT PICKS FOR THE PERIOD (JAN-MAR 2017)

- **Amazon**

Considering that Amazon has been bleeding money in the Asian Subcontinental areas like India due to Flipkart being funded heavily and being pushed out of China due to Alibaba, they're looking to expand widely in areas ranging from Pharmaceuticals to Groceries.

- **Adobe**

Adobe doesn't look to back down as everybody is wishing to move to the digital market and cut down on costs.

- **NVIDIA**

NVIDIA has romped up its sales in competition with AMD, and is releasing Volta GPU's soon which are considered to be next gen. NVIDIA holds an extremely bullish trend as well as high risks, as seen in the wide interval of the Bollinger Bands.

Types of Data and Suitable Charts

Let us see the common types of data we encounter and how to tell stories from those, by selecting the best-fit charts.

Commonly encountered types of data:

1. Textual Data

When data is found in this form, it's usually good to be finding how often a word has been used or what the sentiment of the text is. Stories can be told best using this form of data.

One of the best-suited visualizations for textual data is the WordCloud. The wordcloud brings the more frequent ones to the center and enlarges them, giving us a clear picture of what the general idea of the text depicts.

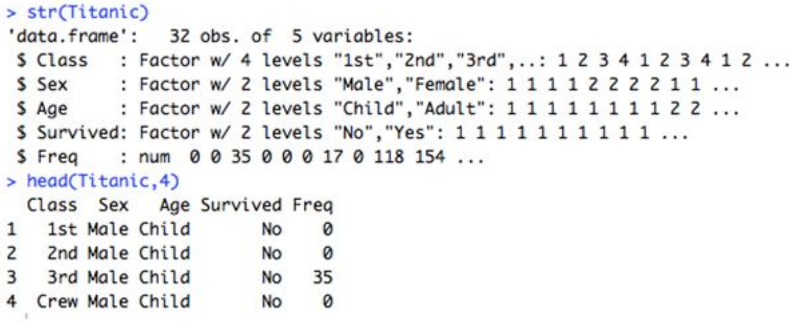
deceit
blame
kill
deep
flaw
ba
haunt
tension
weird
fake
crap
betray
goon
loss
brea
spite
shatt
strike
junk
disappoint
flee
shit
stun
die
twist
mean
reject
plo
fall
overwhelm
low
suc
bore
lack
sad
impromptu
fa
steal
difficult
bleed
disown
stuck
dismal
slap
shock
critic
daze
small
pleas
beat
slow
miser
stole
poor
he
miss

For example, the wordcloud in this article displayed above gives a representation of the twitter dataset. It shows that love is the most frequent positive term used in the tweets.

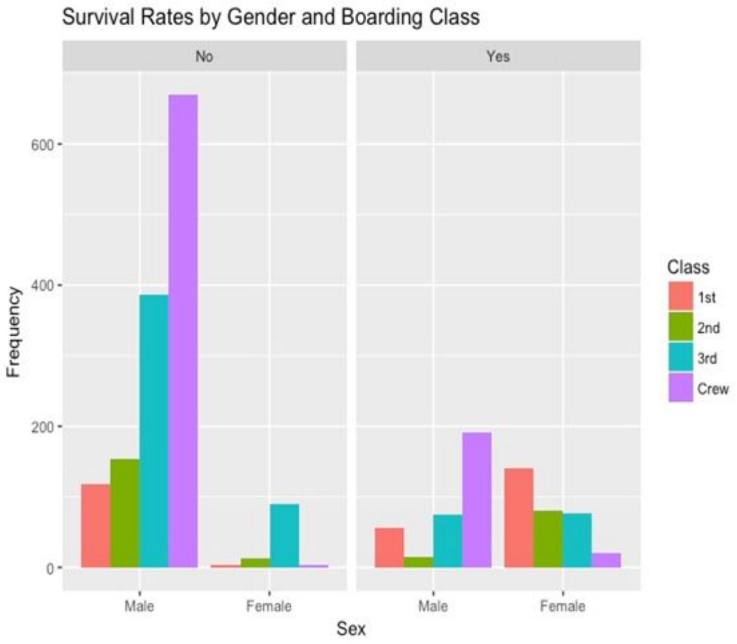
2. Mixed Data

When our data consists of numeric or any other variety of formats, we need to know which ones are important and give us better insights from our dataset.

The preferred visual for this kind of data can vary; here I will show you how to use facet grids for the data. I will be using the Titanic Passenger Data.



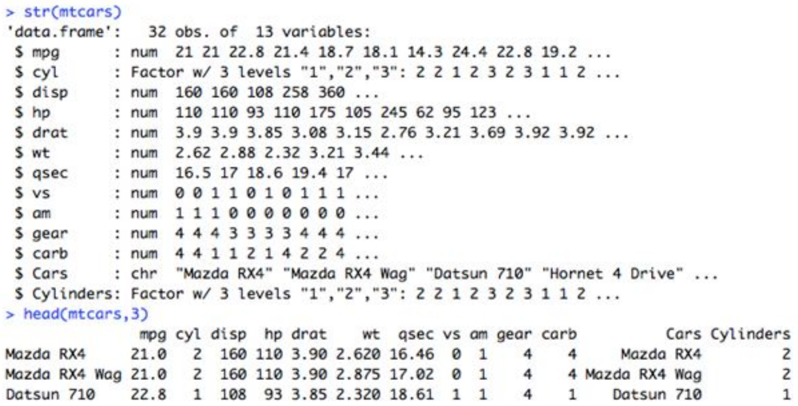
```
> str(Titanic)
'data.frame':   32 obs. of  5 variables:
 $ Class   : Factor w/ 4 levels "1st","2nd","3rd",..: 1 2 3 4 1 2 3 4 1 2 ...
 $ Sex     : Factor w/ 2 levels "Male","Female": 1 1 1 1 2 2 2 2 1 1 ...
 $ Age     : Factor w/ 2 levels "Child","Adult": 1 1 1 1 1 1 1 1 2 2 ...
 $ Survived: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Freq    : num  0 0 35 0 0 0 17 0 118 154 ...
> head(Titanic,4)
  Class  Sex   Age Survived Freq
1   1st Male Child       No    0
2   2nd Male Child       No    0
3   3rd Male Child       No   35
4  Crew Male Child       No    0
```



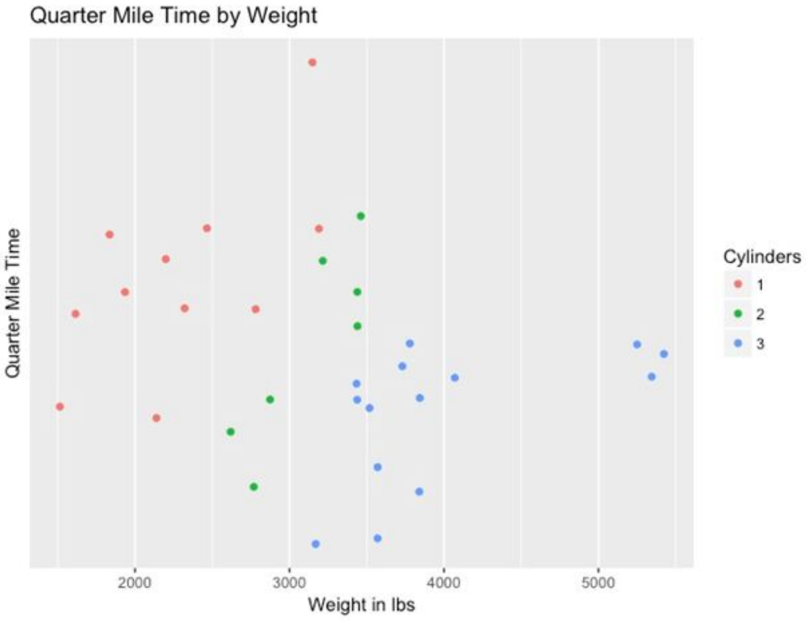
As this plot shows us, females and first-class passengers tend to have a higher survival chance than men who are a part of the crew or lower boarding classes.

Isn’t that what had really happened on the Titanic?

Another way to visualize this kind of data is by trying a multivariate plot. The dataset in use for this plot is the Car Performance and Specifications dataset.



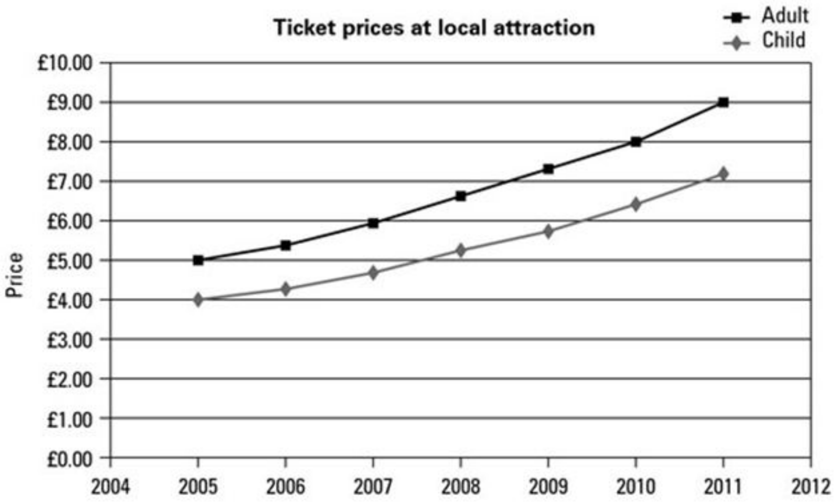
```
> str(mtcars)
'data.frame':	32 obs. of  13 variables:
 $ mpg      : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl      : Factor w/ 3 levels "1","2","3": 2 2 1 2 3 2 3 1 1 2 ...
 $ disp     : num  160 160 108 258 360 ...
 $ hp       : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat     : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt       : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec     : num  16.5 17 18.6 19.4 17 ...
 $ vs       : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am       : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear     : num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb     : num  4 4 1 1 2 1 4 2 2 4 ...
 $ Cars     : chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
 $ Cylinders: Factor w/ 3 levels "1","2","3": 2 2 1 2 3 2 3 1 1 2 ...
> head(mtcars,3)
               mpg cyl disp  hp drat    wt  qsec vs am gear carb          Cars Cylinders
Mazda RX4     21.0   2  160 110 3.90 2.620 16.46  0  1    4    4     Mazda RX4         2
Mazda RX4 Wag 21.0   2  160 110 3.90 2.875 17.02  0  1    4    4 Mazda RX4 Wag         2
Datsun 710    22.8   1  108  93 3.85 2.320 18.61  1  1    4    1    Datsun 710         1
```



Here we can see how Cars that have a heavier built are slower than the ones with lighter bodies. Makes sense, right?

3. Numeric Data

When we encounter this kind of data, we're usually looking for trends or lines that depict numbers. The visual that would suit numeric data best would be a line or a step graph.

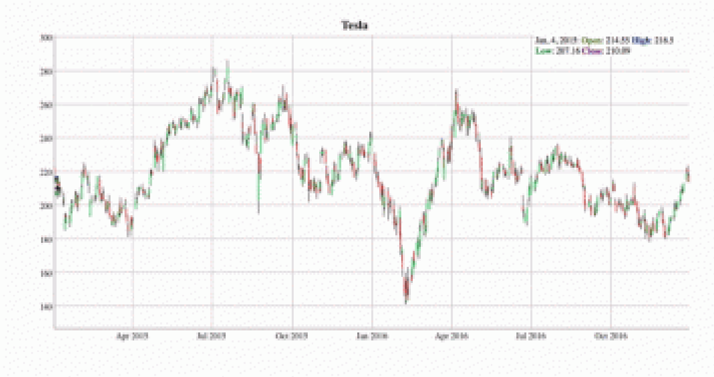


Here, we can very clearly see the rise of prices at a local attraction for adults and children. See how easy it is to see the growth at each year interval?

4. Stocks

One of the datasets that we also encounter are related to stocks. Stock market data is primarily a time series data of numeric values, but as a trader or an investor, I would like to understand each date and drop carefully.

The most visually captivating charts in this regard is the Candlestick chart.

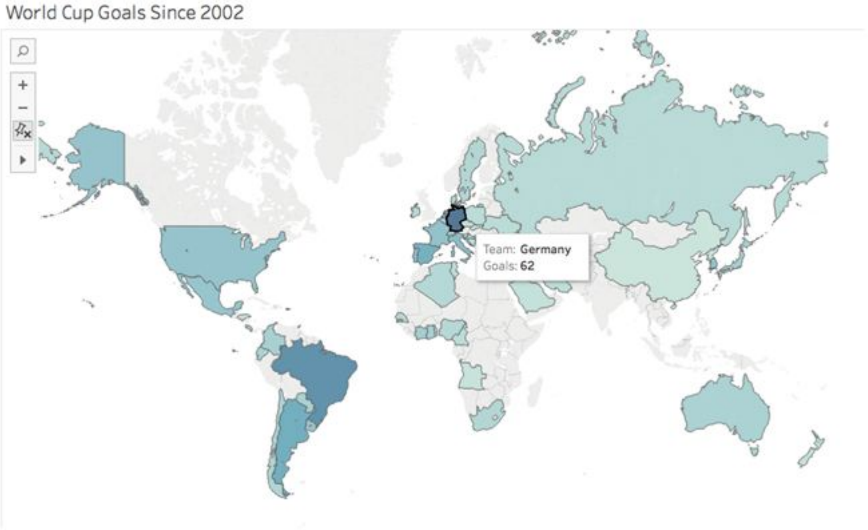


Here, we take the example of Tesla's stocks. The candlestick charts can be used to maneuver across each date and see the lows and highs of stocks individually. This could help us take better investment decisions based on current or past market trends.

As the graph shows us, February 2016 was a drop for Tesla's stocks. We could now use this information to understand other market conditions and economic situations to make decisions about their stock.

5. Geographic Data

When we have data pertaining to specific locations and areas, we use maps to add clarity and meaning to our analysis.



In this example, we can see how countries fared at and after the 2002 World Cup. Germany has scored the maximum number of goals, being one of the most dominant teams in world football ever since.

Storytelling during the steps of predictive modeling

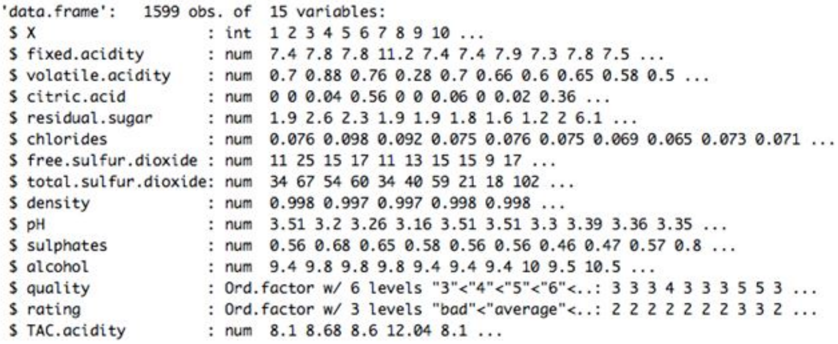
Often, we would be questioned about how our stories and visuals can work or help when it's time to create mathematical models. During all stages of predictive modeling, storytelling could be a vital addition to your analysis.

Let us understand the basic steps involved in creating models out of our data and go through telling stories within them.

1. Data Exploration

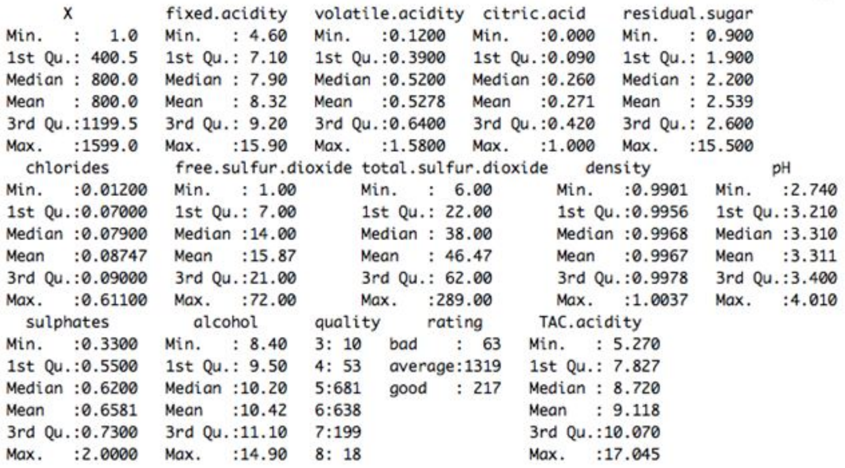
The first step of model building is understanding your data. I'll give you instances and show you how you can explore your data without computing complex statistics.

Let's consider a dataset on Wine Quality. This is the structure of the dataset is as follows



```
'data.frame':   1599 obs. of  15 variables:
 $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3 ...
 $ rating              : Ord.factor w/ 3 levels "bad"<"average"<..: 2 2 2 2 2 2 2 3 3 2 ...
 $ TAC.acidity         : num  8.1 8.68 8.6 12.04 8.1 ...
```

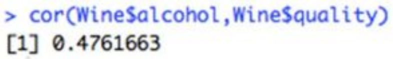
Here, we can see the associated summary statistics of the dataset in use.



```
      X           fixed.acidity   volatile.acidity  citric.acid     residual.sugar
 Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
 1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
 Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
 Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
 3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
 Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
   chlorides        free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   :0.01200   Min.   : 1.00        Min.   :  6.00        Min.   :0.9901   Min.   :2.740
 1st Qu.:0.07000   1st Qu.: 7.00        1st Qu.: 22.00        1st Qu.:0.9956   1st Qu.:3.210
 Median :0.07900   Median :14.00        Median : 38.00        Median :0.9968   Median :3.310
 Mean   :0.08747   Mean   :15.87        Mean   : 46.47        Mean   :0.9967   Mean   :3.311
 3rd Qu.:0.09000   3rd Qu.:21.00        3rd Qu.: 62.00        3rd Qu.:0.9978   3rd Qu.:3.400
 Max.   :0.61100   Max.   :72.00        Max.   :289.00        Max.   :1.0037   Max.   :4.010
   sulphates         alcohol        quality      rating        TAC.acidity
 Min.   :0.3300   Min.   : 8.40   3: 10   bad    :  63   Min.   : 5.270
 1st Qu.:0.5500   1st Qu.: 9.50   4: 53   average:1319   1st Qu.: 7.827
 Median :0.6200   Median :10.20   5:681   good   : 217   Median : 8.720
 Mean   :0.6581   Mean   :10.42   6:638                  Mean   : 9.118
 3rd Qu.:0.7300   3rd Qu.:11.10   7:199                  3rd Qu.:10.070
 Max.   :2.0000   Max.   :14.90   8: 18                  Max.   :17.045
```

So, if we need to see whether there is any correlation between alcohol volumes and wine qualities, how do we do it?

We could either compute Pearson's 'r'. It would help us in building a model, but would not help us in analyzing much.



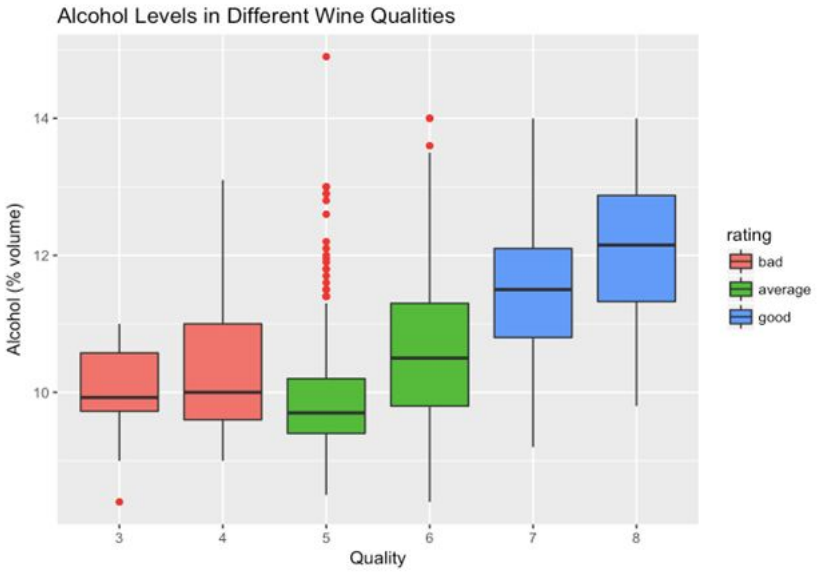
```
> cor(Wine$alcohol,Wine$quality)
[1] 0.4761663
```

This shows a very strong correlation between Alcohol content and wine quality. But does it tell you anything else?

Ideally, it doesn't. So, what does?

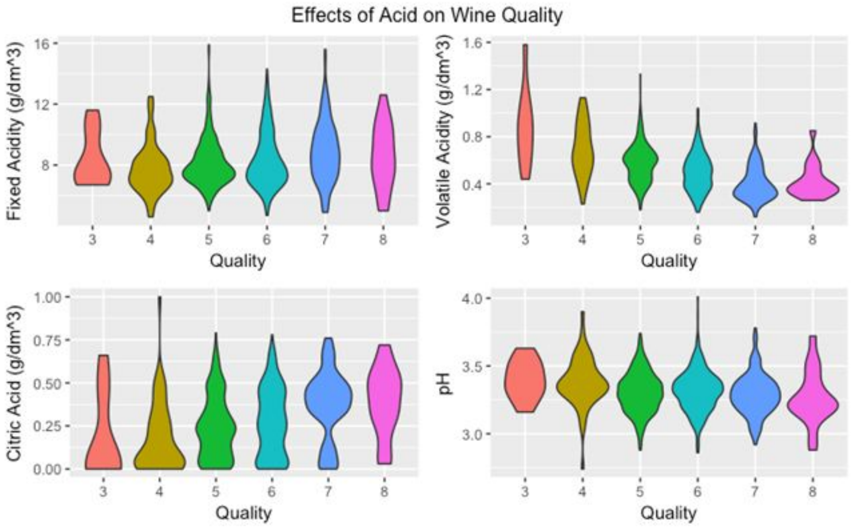
Let's see how we can visualize these and tell a lot more from them.

First, we'll begin by seeing how Wine Quality relates to Alcohol content.



Here, we can see that the higher alcohol volumes relate to better wine qualities and it helps us come to a better understanding of our data. We can also spot outliers better in this scenario.

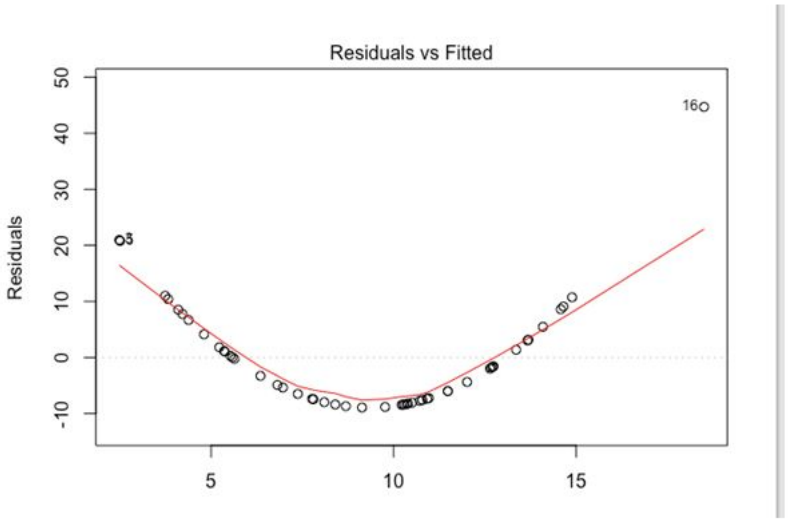
Next, would you wonder how acid contents in your wine affect its quality?



This would be one way to visualize the effects of acid. As the Violin Plot expands horizontally, it shows that there are higher numbers of data points within those areas.

2. Feature Visualizing

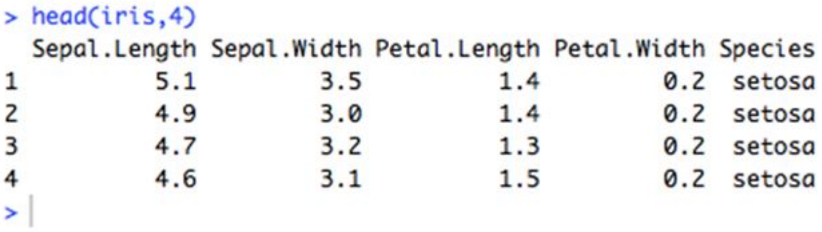
After you generate features, how do you see how well one is predicting?



Graphs tell us how far away our predicted points are from our fitted line.

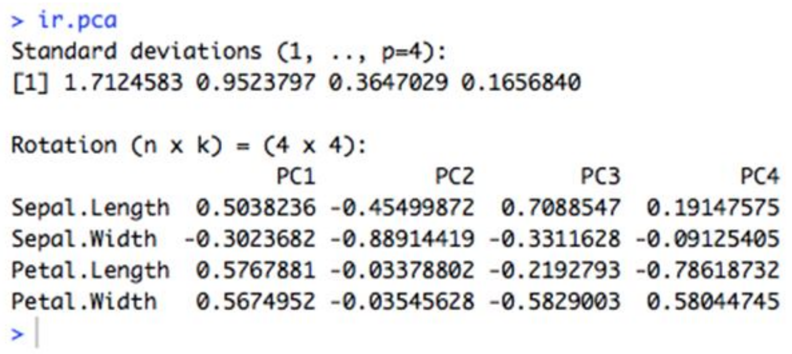
Another example where we might have to visualize newly created visuals is the Principal Component Analysis. If you want to get an in-depth understanding of PCA, you can go through this article.

This is the Iris dataset found in RStudio.



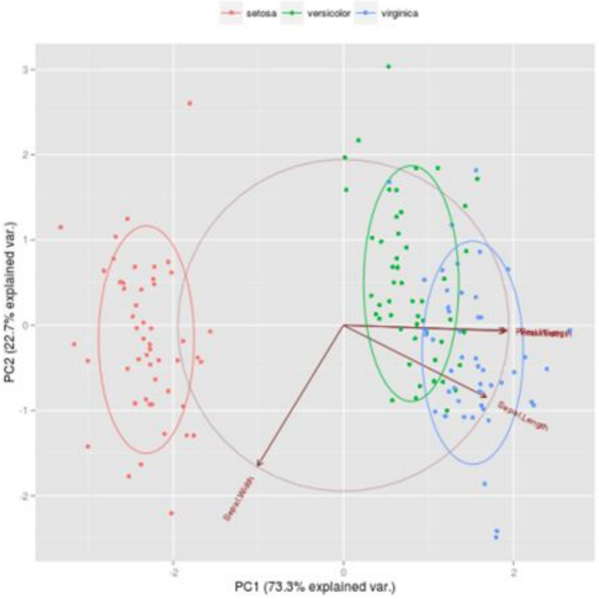
```
> head(iris,4)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
> 
```

When we run the principal component analysis on this dataset, we find these statistics.



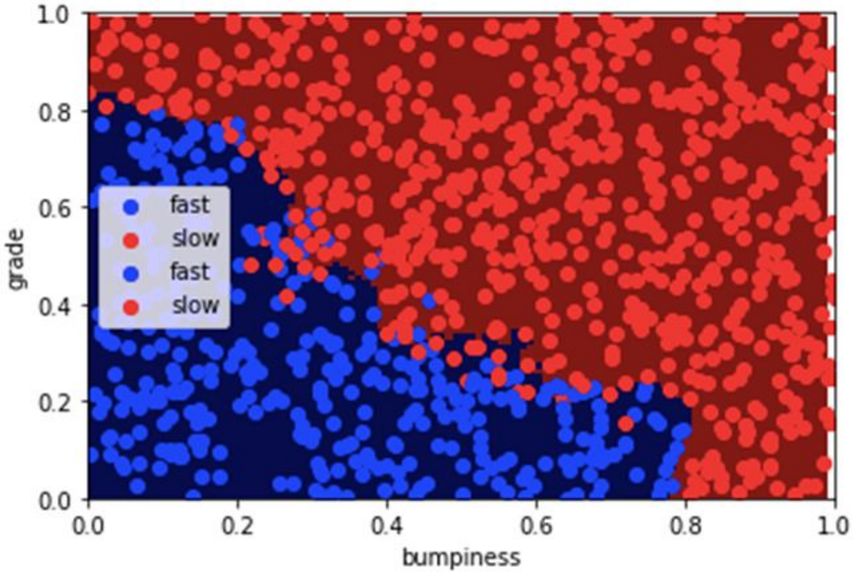
```
> ir.pca
Standard deviations (1, .., p=4):
[1] 1.7124583 0.9523797 0.3647029 0.1656840

Rotation (n x k) = (4 x 4):
                    PC1         PC2        PC3         PC4
Sepal.Length  0.5038236 -0.45499872  0.7088547  0.19147575
Sepal.Width  -0.3023682 -0.88914419 -0.3311628 -0.09125405
Petal.Length  0.5767881 -0.03378802 -0.2192793 -0.78618732
Petal.Width   0.5674952 -0.03545628 -0.5829003  0.58044745
>
```

Although when we plot this, we find that the resulting visual is much more informative than the statistics.



3. Model Creation and Comparison

Coming to the model creation phase, we usually find the need to understand how our data is being fitted.

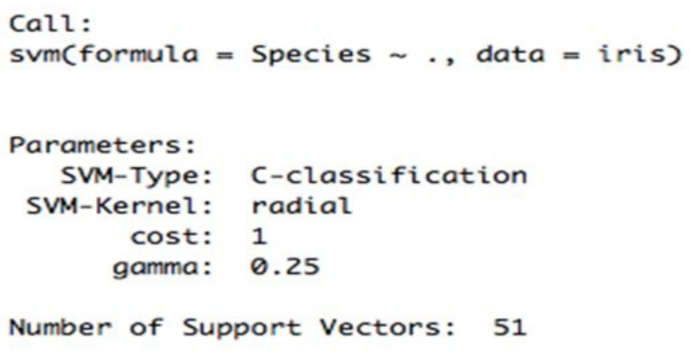


This is a model that predicts whether the car should go fast or slow, based on the grade of the road and bumpiness.

As you can see, the decision boundary clearly classifies most of the data but an accuracy of 88.21% doesn't tell much of a story. Here we can even see how far the misclassified points are from the decision boundary.

We can also compare certain algorithms and techniques by looking at their decision boundaries as we did above.

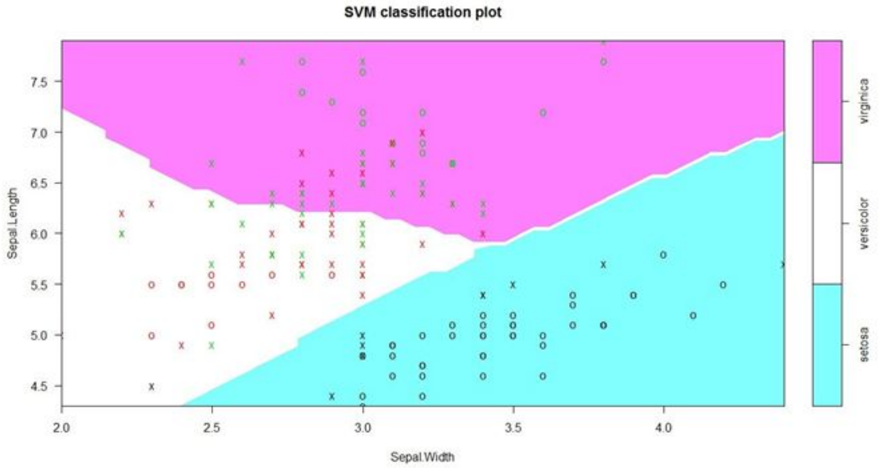
Another example using the Iris dataset is shown below.



```
Call:
svm(formula = Species ~ ., data = iris)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.25

Number of Support Vectors:  51
```

Here, there's not much information to derive valuable insights about our model.

To learn more about Support Vector Machines, you can go through [this article](#).



On the other hand, this plot shows us a clear classification boundary where the Species separate from each other.

Best Practices for Story Telling

Now that you know the scenarios where we can use story telling to explain our point, I will give you a few practical tips when you take this up on your own.

- Always **label your axes and give the heading** of your plot.
- Use **legends** where necessary.

- Use **colors that are lighter** on the eye and in proportion.
- **Avoid adding unnecessary detail** to your visualization like backgrounds or themes that don't allow good readability.
- **Only a point can be used** to simultaneously encode two quantitative values based on a horizontal and vertical location.
- **Never use points** for visualization if you are doing time series encoding.

End Notes

Storytelling is more than what it has been used for. It can uncover insights from your data that you might have missed before. Relations between features and data that numbers can never clearly depict, can be shown using stories and charts.

In this piece, we've elaborated on how stories are used in almost all avenues to explain a detail better. Starting from how they're used in the steps of model building, we've gradually gone on to which charts suit specific data types well.

I hope you had a great time reading the article. Eager to hear your data stories!

Article Url - https://www.analyticsvidhya.com/blog/2020/05/art-storytelling-analytics-data-science/

Analytics Vidhya

This is the official account of the Analytics Vidhya team.