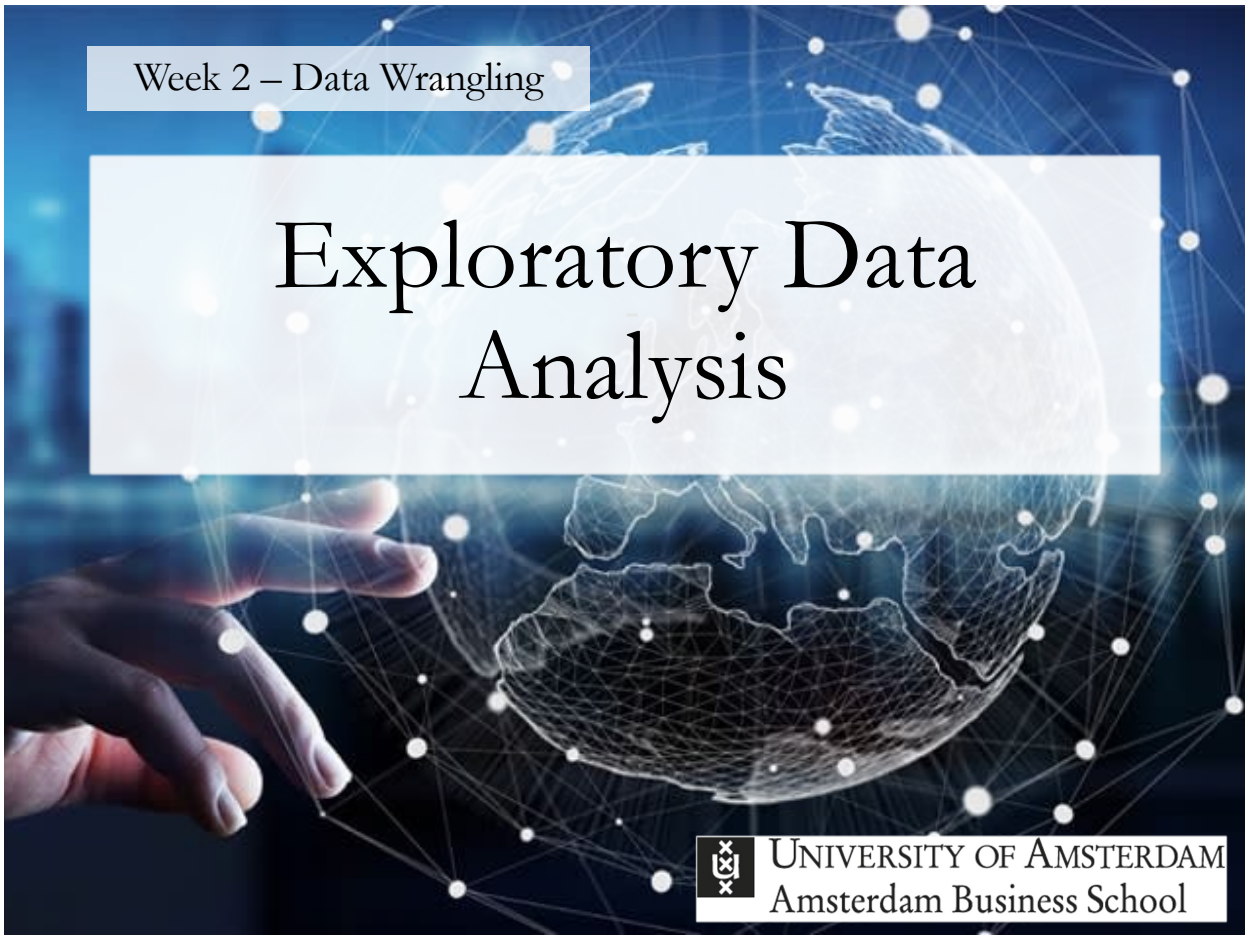


Contents

Recapture: Data Science

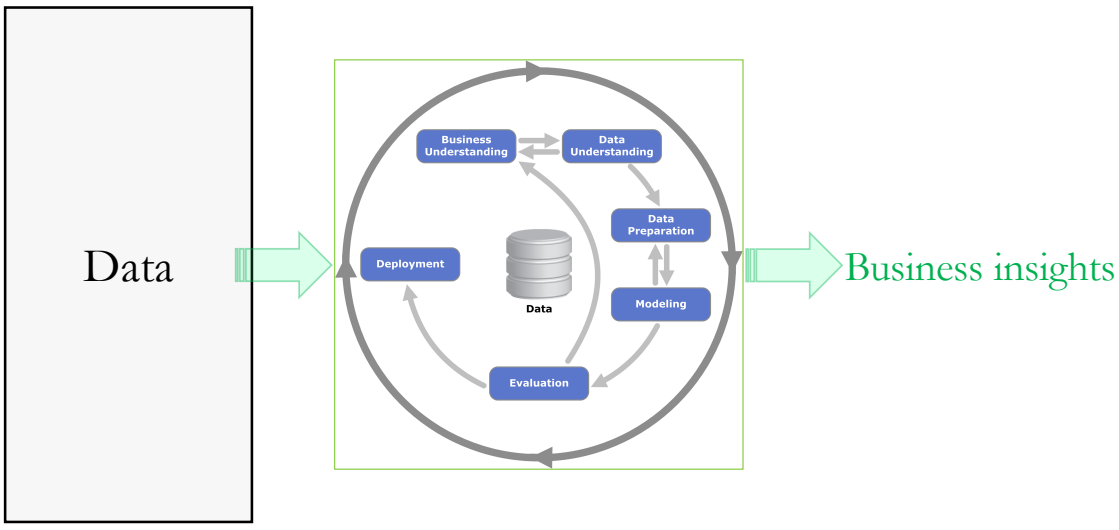
Data Science uses tools and techniques to turn data into meaningful business insights.

Goal: Use data to take better business decisions.

Recapture: Data Science

Data Science uses tools and techniques to turn data into meaningful business insights.

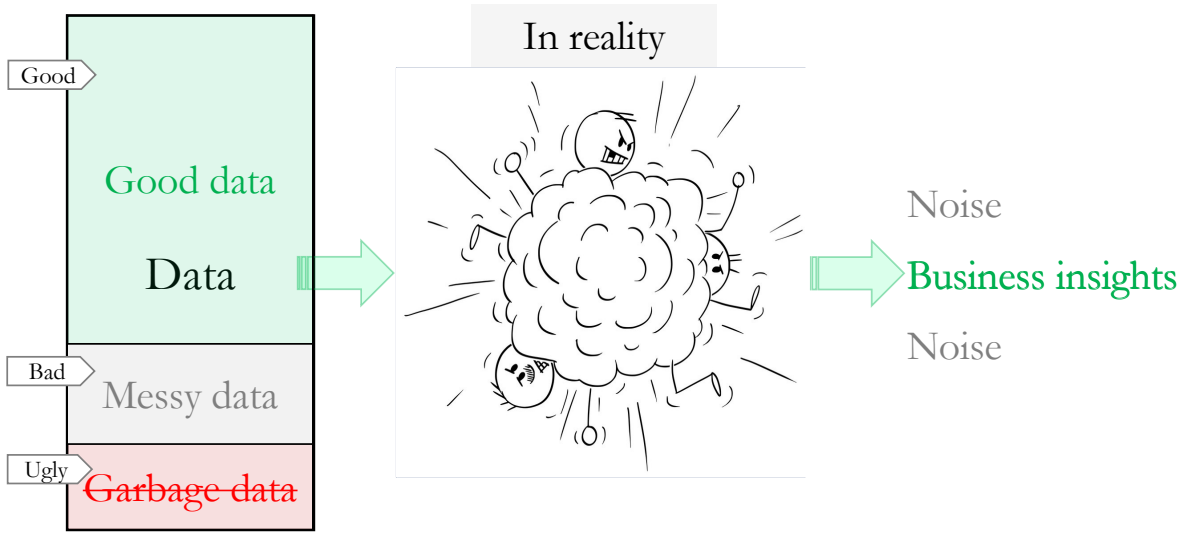
Goal: Use data to take better business decisions.



Recapture: Data Science

Data Science uses tools and techniques to turn data into meaningful business insights.

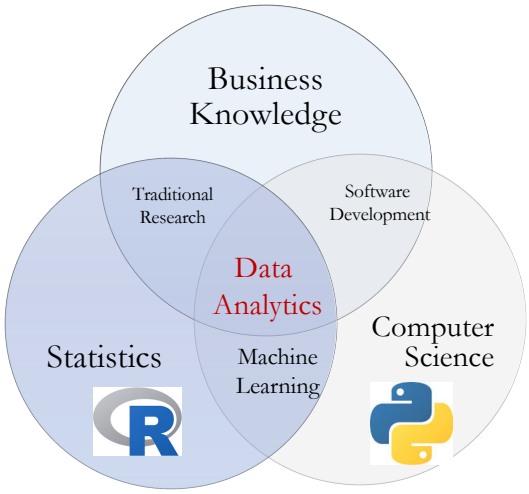
Goal: Use data to take better business decisions.



Recapture: Data Science

A lot of different terms are around analytics likes:

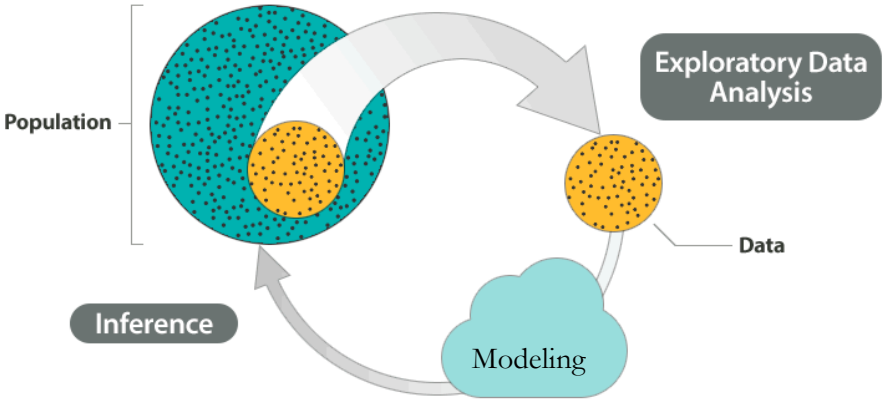
- Data Analytics
- Business Intelligence
- Data Science
- Decision Science
- Data Mining
- ….



Although there are some differences, all of them support the same goal: "turning data into useful insight".

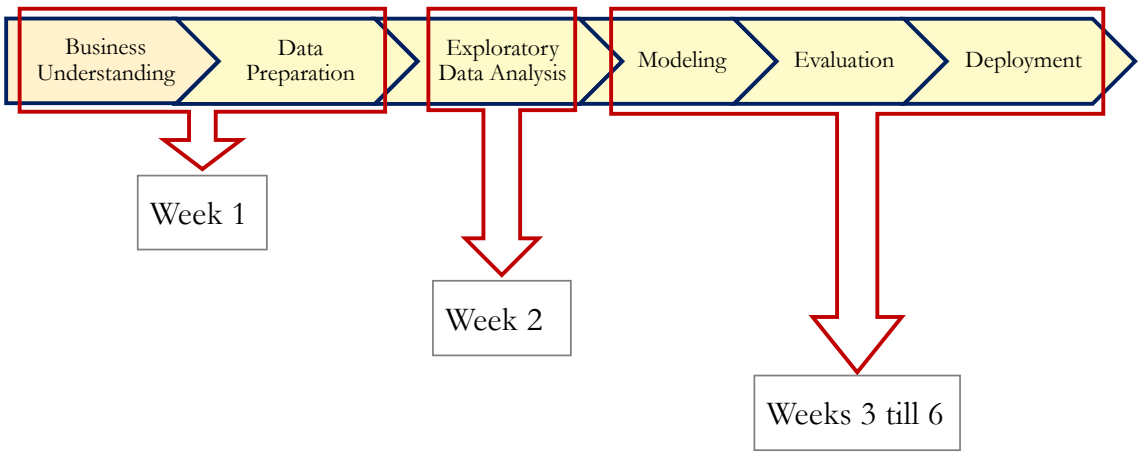
Recapture: 6-Step of Data Science

6-Step of Data Science process is a general problem-solving strategy of business/research unit, which is adaptive life cycle

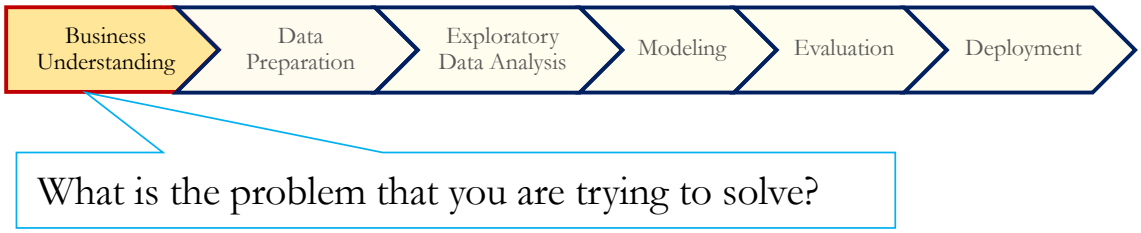


Recapture: 6-Step of Data Science

6-Step of Data Science process is a general problem-solving strategy of business/research unit, which is adaptive life cycle



Recapture: Business Understanding



In Business understanding phase we basically:

1. Understands the business process;
2. Define project requirements and objectives;
3. Translate objectives into a data analytics problem definition;
4. Prepare a preliminary strategy to meet objectives.

Recapture: Business Understanding

In the exercises of week 1, your task was to write a Business problem that you have (or you had) in your company by following the three steps:

1. First, clearly enunciate the project objectives and requirements in terms of the business unit as a whole.
2. Then, translate these goals and restrictions into the formulation of a problem that can be solved using data science.
3. Finally, prepare a preliminary strategy for achieving these objectives.

Each group should represent the result in the class for around 5 minutes.

Recapture: Business Understanding

Example: Tech Company

Intl.Mins	Intl.Calls	Intl.Charge	CustServ.Calls	Churn
10.0	3	2.70	1	False
13.7	3	3.70	1	False
12.2	5	3.29	0	False
6.6	7	1.78	2	False
10.1	3	2.73	3	False
6.3	6	1.70	0	False
7.5	7	2.03	3	False
7.1	6	1.92	0	False
8.7	4	2.35	1	False
11.2	5	3.02	0	False
12.7	6	3.43	4	True
9.1	5	2.46	0	False
11.2	2	3.02	1	False

Churn Data

Business insights:

Predict behavior to retain customers.

This dataset is availed in the R package "liver":

Recapture: Data Preparation

Business Understanding → Data Preparation → Exploratory Data Analysis → Modeling → Evaluation → Deployment

- Clean and prepare data so it is ready for modeling tools.
- Perform transformation of certain variables, if needed.

Raw data are often unprocessed, incomplete, noisy and may contain:

- Obsolete/redundant fields
- Missing values
- Outliers
- Data in a form not suitable for data analysis
- Values not consistent with policy or common sense

Recapture: Data Preparation

For data analytics purposes, database values must undergo data cleaning and data transformation.

Minimize GIGO (Garbage In → Garbage Out).

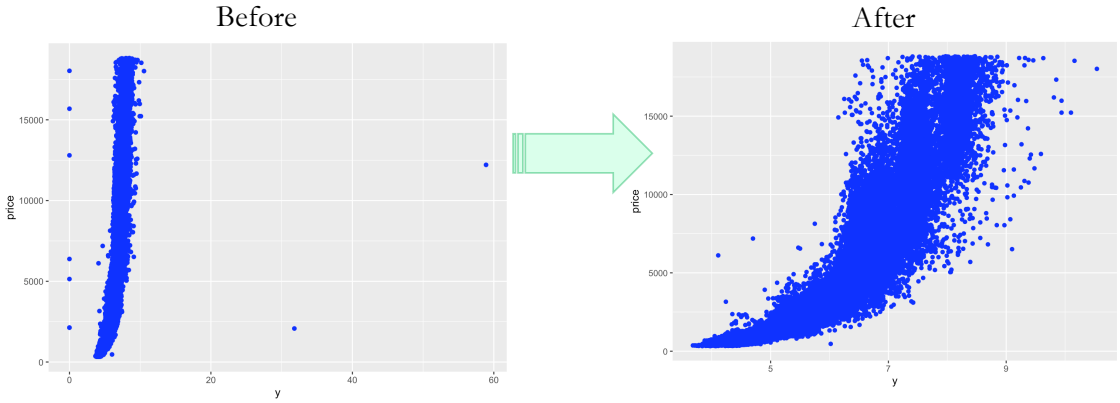
- IF GIGO is minimized → THEN Garbage results Out from model is minimized.

Effort for data preparation ranges around 10%-60% of data analysis process – depending on the dataset.

Recapture: Data Preparation

Data Cleaning:

Detecting unusual values in the Diamonds Dataset

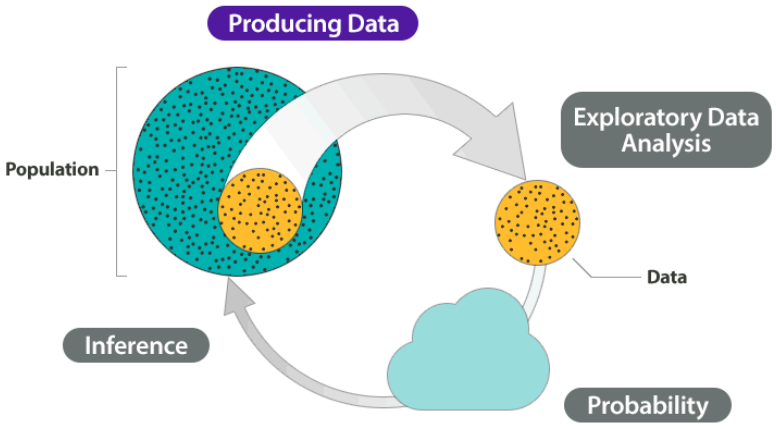


Contents

Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyzing data to summarize their main characteristics, often with visual methods. e.g.

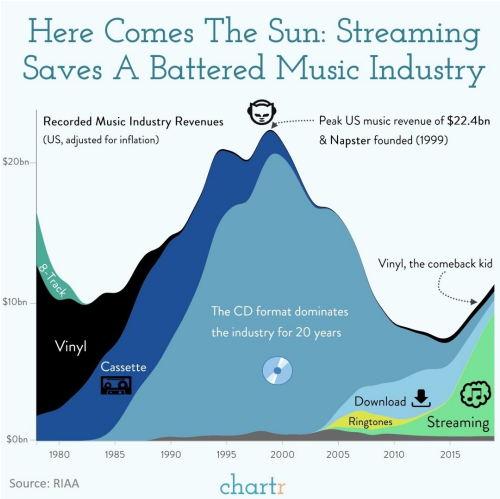
- How coronavirus affects the global financial system.
- Examining important inter-relationships between attributes.
- Developing an initial idea of possible associations amongst the predictors, as well as between the predictors and the target variable.

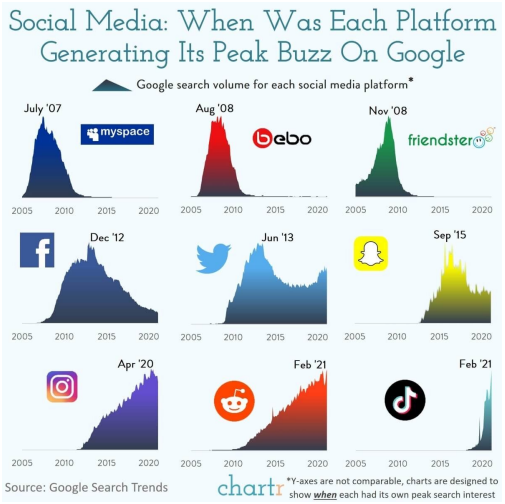


Exploratory Data Analysis

Exploratory Data Analysis and Data Storytelling

https://www.chartr.co/



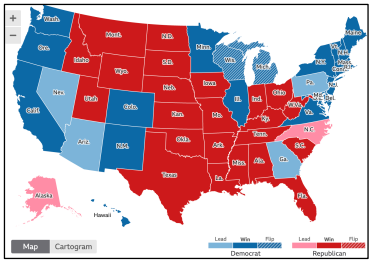


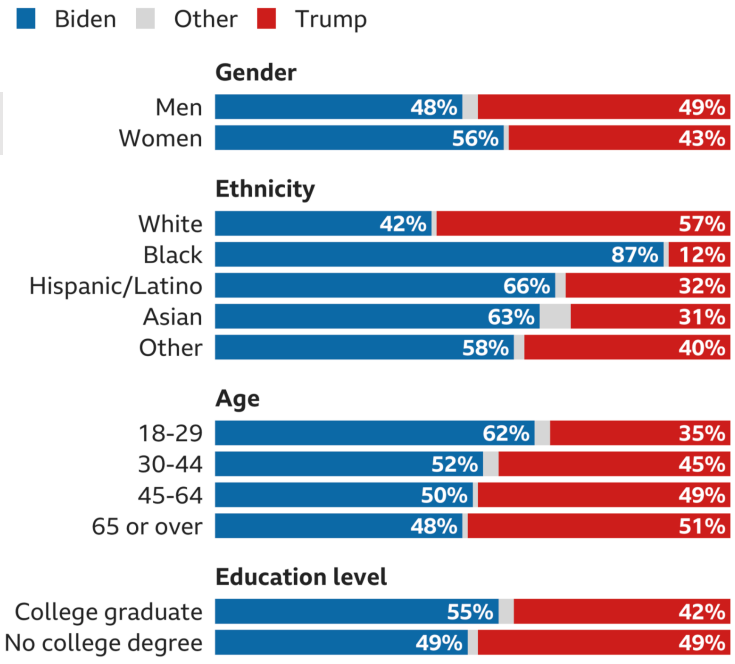
Source: BBC news

Exploratory Data Analysis

Example:

US election in 2020



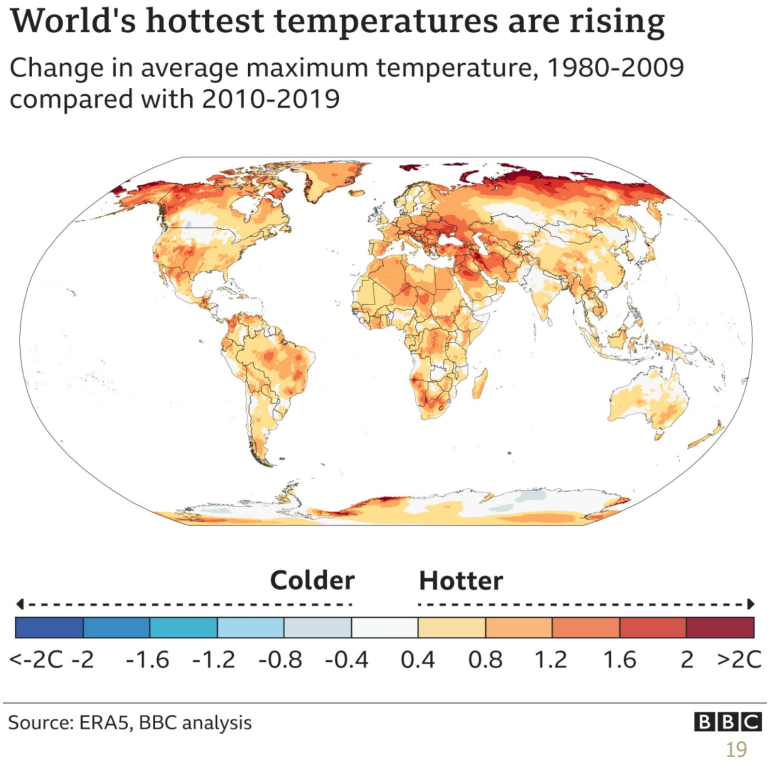


Source: BBC news

Exploratory Data Analysis

Example:

Climate change: World now sees twice as many days over 50C.



Exploratory Data Analysis

Ted Talk: The beauty of data visualization

https://youtu.be/5Zg-C8AAIGg

EDA based on Churn Dataset

Following slides is an EDA based on Churn data set.

- Churn data set contains 5000 records (customers) and 20 variables.
- Churn indicates a customer leaving a company's service in favor of another.

Intl.Mins	Intl.Calls	Intl.Charge	CustServ.Calls	Churn
10.0	3	2.70	1	False
13.7	3	3.70	1	False
12.2	5	3.29	0	False
6.6	7	1.78	2	False
10.1	3	2.73	3	False
6.3	6	1.70	0	False
7.5	7	2.03	3	False
7.1	6	1.92	0	False
8.7	4	2.35	1	False
11.2	5	3.02	0	False
12.7	6	3.43	4	True
9.1	5	2.46	0	False
11.2	2	3.02	1	False

Churn Data

Business insights:

Predict behavior to retain customers.

Business Understanding

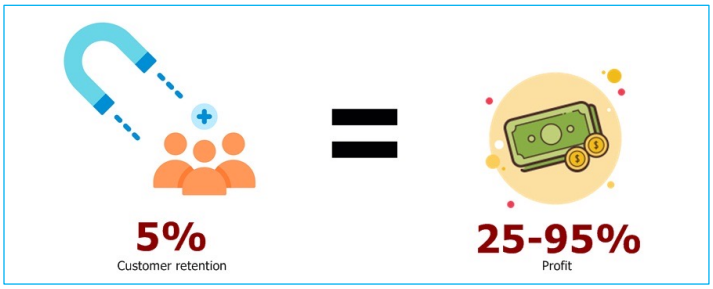
Customer Churn (Attrition) occurs when customers stop doing business with a company.

- Customer retention is essential for future profit maximization.
- A good retention strategy should answer both the Who and the When questions.
- The key to any retention strategy is to predict who is at risk at churning and how valuable they are to the company. So, we can decide whether to take action.
- If we decide to take corrective measures, predictive analytics will help us decide what they should be by identifying the causes of our customer's churn.

Business Understanding

Customer Churn & Profitability:

On average, 5% reduction in churn rates translates to 25% to 95% increase in profit.



- True cost of losing customers is not fully considered – (5 times cheaper to retain a customer than acquire a new one).
- With 80% of business coming from 20% of customers, repeat customers are key to survival!
- Because churn is often poorly defined and tracked, retention efforts end up being costly or ineffective.

Getting to Know the Data Set

```
str( churn )   # Compactly display the structure of the data
```

```
## 'data.frame':    5000 obs. of  20 variables:
##  $ state          : Factor w/ 51 levels "AK","AL","AR",..: 17 36 32 36 37 2 20 25 19 50 ...
##  $ area.code      : Factor w/ 3 levels "area_code_408",..: 2 2 2 1 2 3 3 2 1 2 ...
##  $ account.length: int  128 107 137 84 75 118 121 147 117 141 ...
##  $ voice.plan     : Factor w/ 2 levels "yes","no": 1 1 2 2 2 2 1 2 2 1 ...
##  $ voice.messages: int  25 26 0 0 0 0 24 0 0 37 ...
##  $ intl.plan      : Factor w/ 2 levels "yes","no": 2 2 2 1 1 1 2 1 2 1 ...
##  $ intl.mins      : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
##  $ intl.calls     : int  3 3 5 7 3 6 7 6 4 5 ...
##  $ intl.charge    : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
##  $ day.mins       : num  265 162 243 299 167 ...
##  $ day.calls      : int  110 123 114 71 113 98 88 79 97 84 ...
##  $ day.charge     : num  45.1 27.5 41.4 50.9 28.3 ...
##  $ eve.mins       : num  197.4 195.5 121.2 61.9 148.3 ...
##  $ eve.calls      : int  99 103 110 88 122 101 108 94 80 111 ...
##  $ eve.charge     : num  16.78 16.62 10.3 5.26 12.61 ...
##  $ night.mins     : num  245 254 163 197 187 ...
##  $ night.calls    : int  91 103 104 89 121 118 118 96 90 97 ...
##  $ night.charge   : num  11.01 11.45 7.32 8.86 8.41 ...
##  $ customer.calls: int  1 1 0 2 3 0 3 0 1 0 ...
##  $ churn          : Factor w/ 2 levels "yes","no": 2 2 2 2 2 2 2 2 2 2 ...
```

Getting to Know the Data Set

1. State: Categorical, for the 50 states and the District of Columbia,
2. Account Length: Integer-valued, how long account has been active,
3. Area code: Categorical,
4. International Plan: Dichotomous categorical, yes or no,
5. Voice Mail Plan, Dichotomous categorical, yes or no,
6. Number of Voice Mail Messages: Integer-valued,
7. Total Day Minutes: Continuous, minutes customer used service during the day,
8. Total Day Calls: Integer-valued,
9. Total Day Charge: Continuous, perhaps based on above two variables,
10. Total Eve Minutes: Continuous, minutes customer used service during the evening,
11. Total Eve Calls: Integer-valued,
12. Total Eve Charge: Continuous, perhaps based on above two variables,
13. Total Night Minutes: Continuous, minutes customer used service during the night,
14. Total Night Calls: Integer-valued,
15. Total Night Charge: Continuous, perhaps based on above two variables,
16. Total International Minutes: Continuous, minutes customer used service to make international calls,
17. Total International Calls: Integer-valued,
18. Total International Charge: Continuous, perhaps based on above two variables,
19. Number of Calls to Customer Service: Integer-valued,
20. Churn: Target. Indicator of whether the customer has left the company (yes or no).

What type of variables we have?

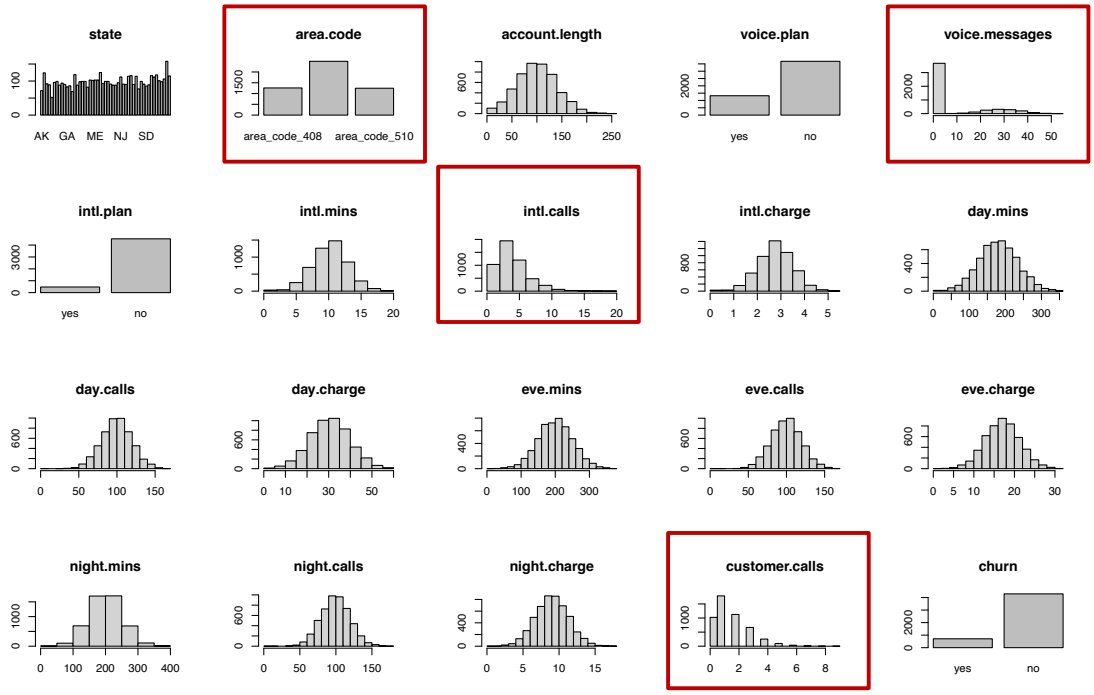
➢Numerical:

- Continuous (entities get a distinct score), e.g. temperature, body length.
- Discrete (counts), e.g.: number of defects.

➢Categorical (entities are divided into distinct categories):

- Binary variable (two outcomes), e.g. dead or alive.
- Ordinal variable, e.g. bad, intermediate, good.
- Nominal variable, e.g. whether someone is an omnivore, vegetarian or vegan.

Getting to Know the Data Set



Getting to Know the Data Set

Insights:

- Vmail messages has a spike on the length.
- Most quantitative variables seem normally distributed, except for "intl.calls" and "customer.calls", which are right-skewed.
- It shows 51 for State, but only 3 for Area Code – how can this be?

Contents

Exploratory Data Analysis

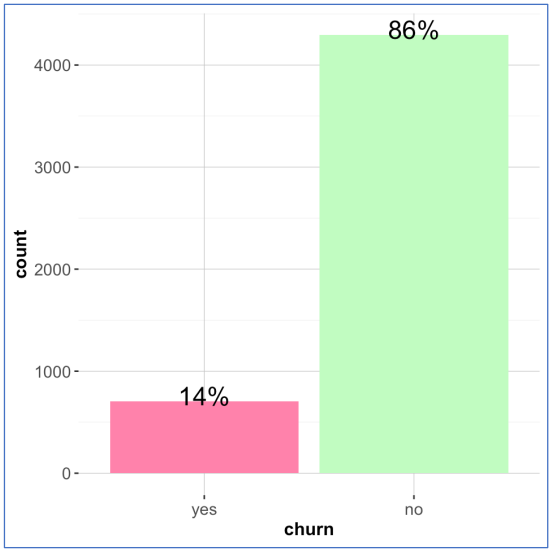
Investigate variables

- Examine Distributions of categorical variables,
- Look at Histograms of numerical variables,
- Explore relationships among sets of variables.

Specific goal for Churn data mining example is to develop a model for the type of customer likely to churn.

Objective: Explore the data while keeping an eye on the overall picture.

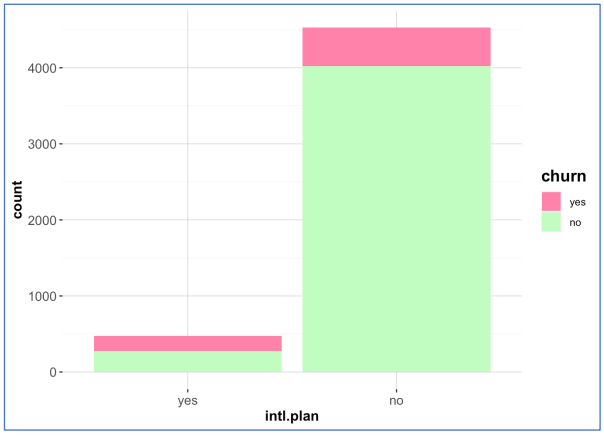
Target Variable



From those 5000 customers 14% of them they left the company

International Plan

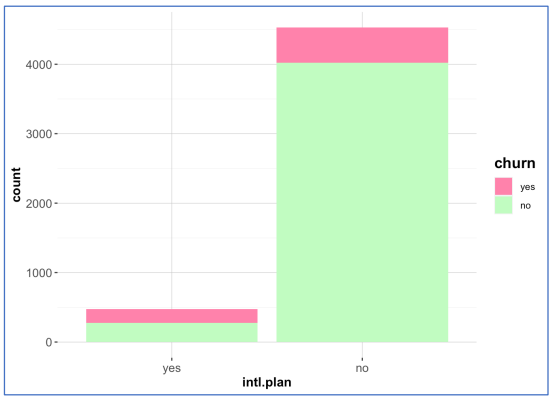
The figure below shows the proportion of customers in International Plan with a churn overlay.

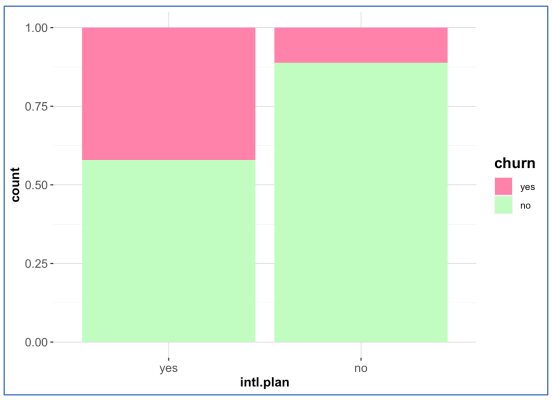


Possibly, greater proportion of those in International Plan are churners?
Difficult to confirm, since the two bars are at a different scale.

International Plan

Same sized bars: Now, same-sized bars used for each category. This allows to compare proportions among different categories.





Those selecting International Plan are more likely to churn. But relationship is not quantified.

International Plan

Contingency table is to quantify the relationship between two categorical variables.

Example: Contingency table for both categorical variables International Plan and Churn.

		International Plan		
		No	Yes	Total
Churn	False	2664	186	2850
	True	346	137	483
	Total	3010	323	3333

Adds up all customers who didn't churn.

Adds up all customers who did churn.

Adds up all customers, regardless of Churn. This is the marginal distribution for International Plan.

Adds up all customers without International Plan.

Adds up all customers with International Plan.

Adds up all customers, regardless of International Plan.
This called the marginal distribution for Churn, shows the frequency distribution for the variable Churn.

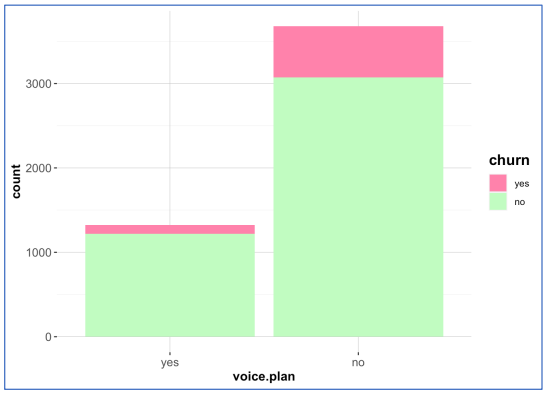
International Plan

Summary of the EDA for International Plan

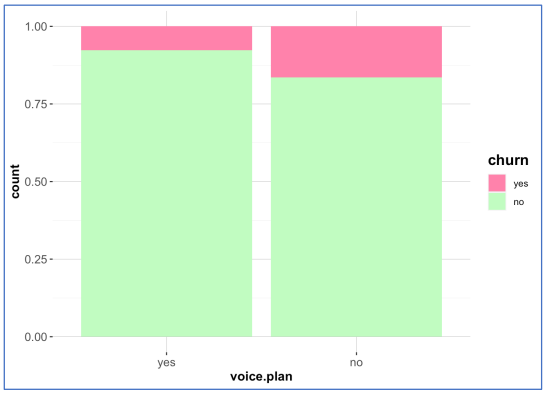
- Perhaps we should investigate what it is about our international plan that is inducing our customers to leave.
- We should expect that, whatever data mining algorithms we use to predict churn, the model will probably include the variable International Plan.

Voice Mail Plan

Proportion of customers in the Voice Mail Plan (bars) with churn overlay.



Same-sized bars for the Proportion of customers in the Voice Mail Plan with churn overlay (normalized).



Those who not participating in the Voice Mail Plan appear more likely to churn.

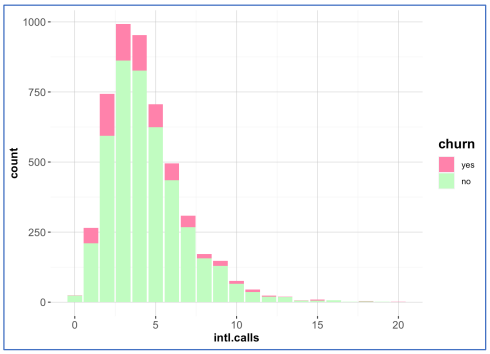
Voice Mail Plan

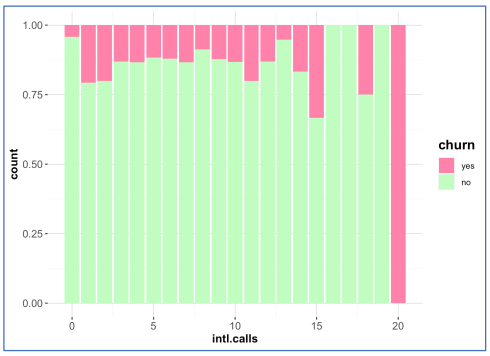
Summary of the EDA for Voice Mail Plan

- Consider enhancing the Voice Mail Plan or offer, as it seems to increase customer loyalty.
- Expect algorithms to probably include the variable Voice Mail Plan. This expectation is not quite as strong as the one found for the International Plan.

Contents

Customer Service Calls





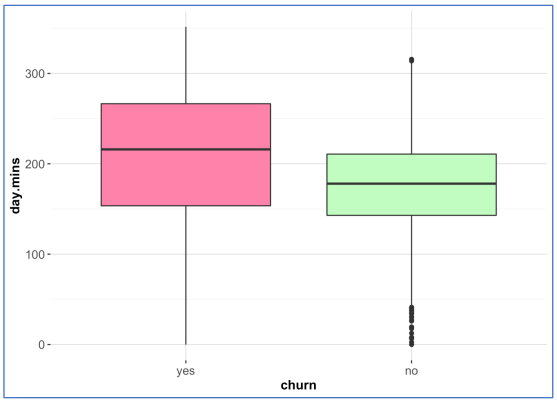
Histogram for Customer Service Calls including Churn overlay.

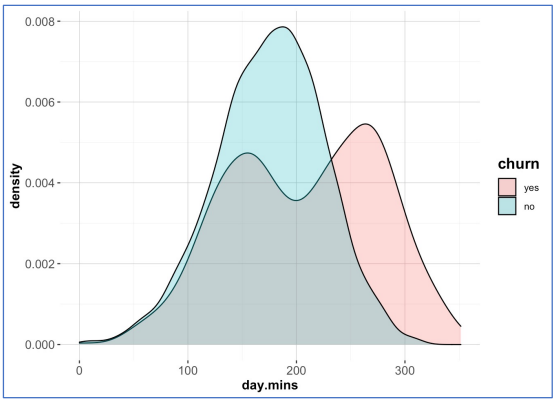
Determining whether Churn proportion varies across number of Customer Service Calls is difficult to recognize.

Proportional histogram of Customer Service Calls - it shows that customers who call customer service more than 3 times are more likely to churn.

Day Minutes

Since the variable "*Day Minutes*" is numerical, we report the boxplot and density:

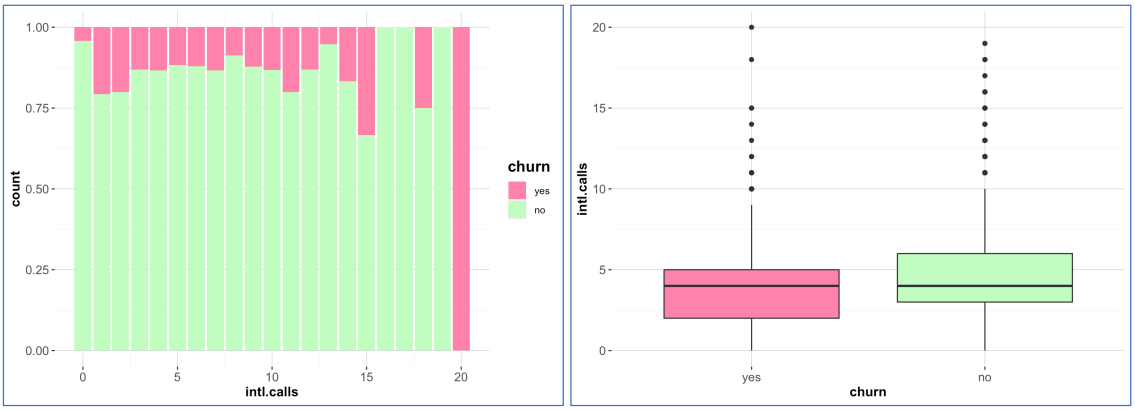




- We should carefully track customer Day Minutes as total exceeds 200. Investigate why those with high usage tend to leave.
- We should consider using variable "*Day Minutes*" as one of the predictor variables in whatever machine learning algorithms we use to predict churn.

International Calls

Since variable "International Calls" is discreate and variable churn is categorical variable, we could us histogram and boxplot :



The above plots do not indicate strong graphical evidence of the predictive importance of International Calls.

Exploring Numeric Variables

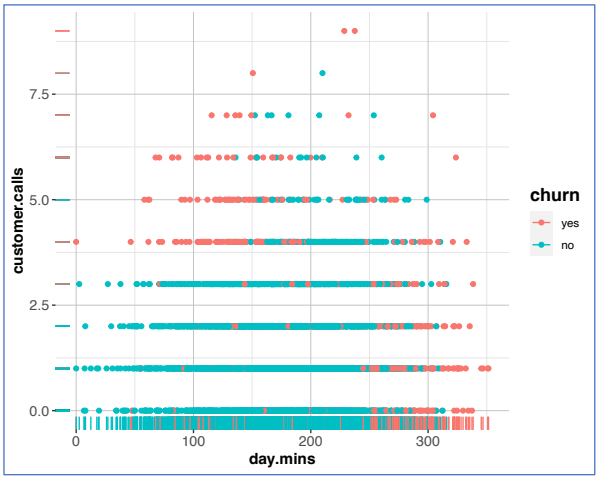
- Additional EDA concludes no obvious association between Churn and the remaining numeric variables (not shown). These numeric variables are probably not strong predictors in the data model.
- However, they should be retained as input to the model. Important higher-level associations/interactions may exist. In this case, let the model identify which inputs are important.
- Different EDA tasks may involve larger numbers of inputs.
- The data analysis performance is adversely affected by the number of inputs.
- If possible, exclude inputs not associated with the target variable.
- Or, use dimension-reduction techniques such as principal components analysis to reduce the number of inputs.

Contents

Exploring Multivariate Relationships

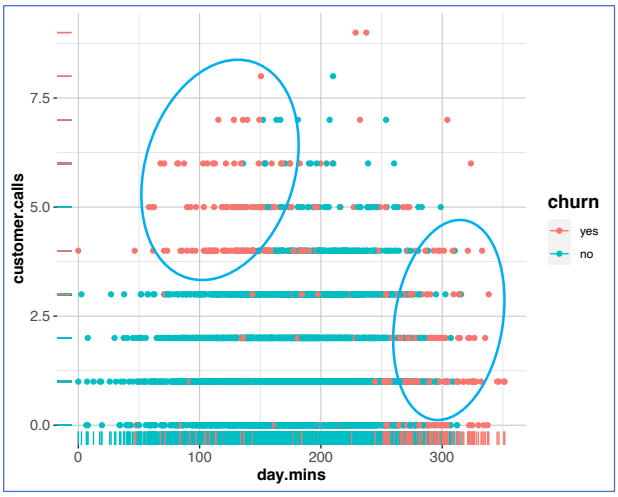
- Possible multivariate relationships examined.
- Two and three-dimensional scatter plots used.

Example: The figure below shows a scatter plot of Customer Service Calls versus Day Minutes.



- Upper-left quadrant indicates high-churn area.
- Identifies customers with high number of customer service calls, combined with low day minute usage.

Exploring Multivariate Relationships

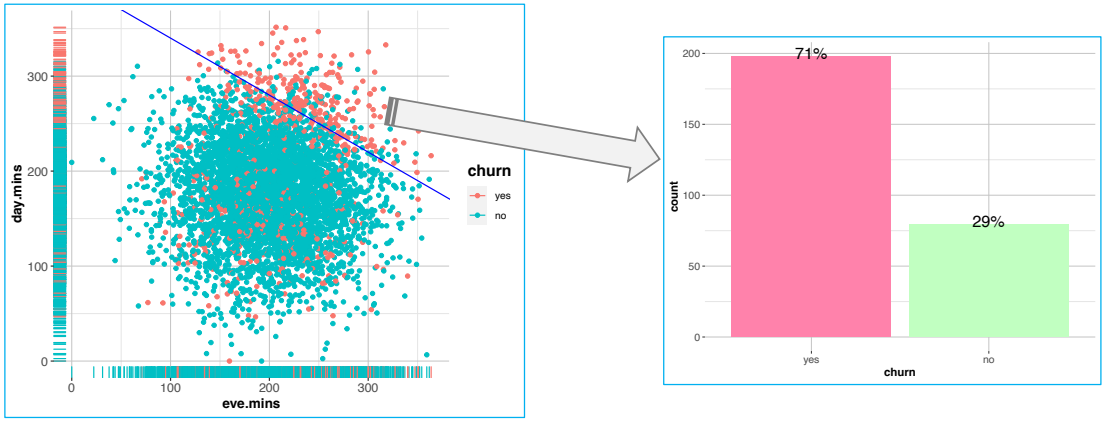


- This relationship not detected using univariate analysis.
- Note, interaction between two variables makes association apparent.
- Univariate analysis determined customers with high number Customer Service Calls churn at higher rates.

Selecting Subsets of the Data

Scatter plots or histograms identify interesting subsets of data.

Example: Left figure shows selection of churners with high Day and Evening Minutes and right figure shows distribution of Churn for this subset.



71% of customers having both high day and evening minutes are churners that is around 5 times churn rate of entire data set.

Contents

Dealing with Correlated Variables

Correlated variables in the model:

- Should be avoided!
- Incorrectly emphasizes one or more data inputs.
- Creates model instability and produces unreliable results.

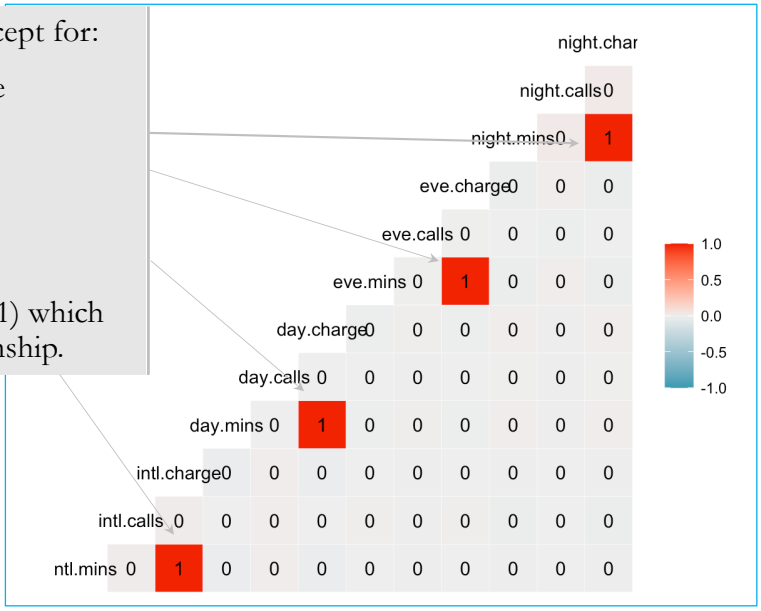
Exploring Numeric Variables

Recall that the use of correlated variables should be avoided.

All correlations are "Weak" except for:

- night.mins and night.charge
- eve.mins and eve.charge
- day.mins and day.charge
- intl.mins and intl.charge

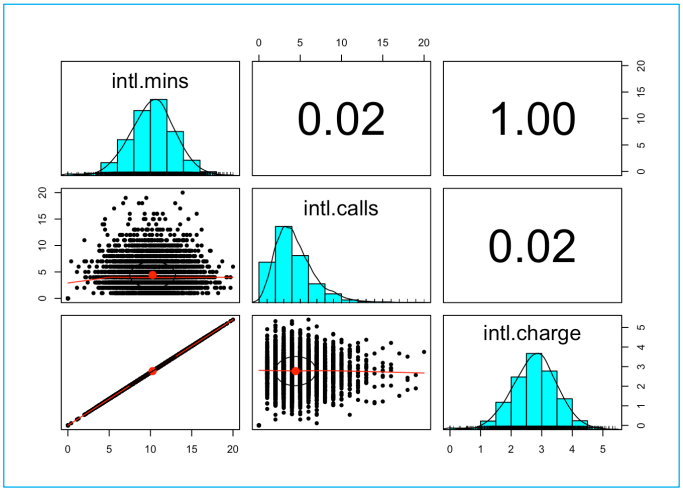
they have high correlation (r=1) which indicate a perfect linear relationship.



What would be your conclusion?

Dealing with Correlated Variables

- As the number of International Minutes increase, we expect International Charge to increase; Example of positive correlation.
- a linear relationship between International Charge & International Minutes.
- International Charge is a linear function of International Minutes.
- Additionally, r = 0.02 indicating that the variables are uncorrelated.



Dealing with Correlated Variables

- One of the two variables should be eliminated from the model.
- Day Charge is arbitrarily chosen for removal.
- Evening, Night and International variable pairs reflect similar results.
- Therefore, Evening Charge, Night Charge and International Charge are also removed.
- Proceeding to data mining without first eliminating these correlated variables may have produced compromised results.
- The number of variables is reduced from 20 to 16.
- Reduction in dimensionality of solution space is beneficial to some data mining algorithms.

Summary

EDA uncovered some insights into Churn data set:

- The four "Charge" variables are linear functions of "Minutes" variables.
- Correlation among remaining numeric attributes is "Weak".
- Area Code and/or State fields are anomalous.
- Customers with International Plan churn at higher rate.
- Customers in Voice Mail Plan churn less frequently.
- Customers calling customer service 4 or more times churn ~4X higher than others.
- Customers with high day and evening minutes churn ~6X higher than others.

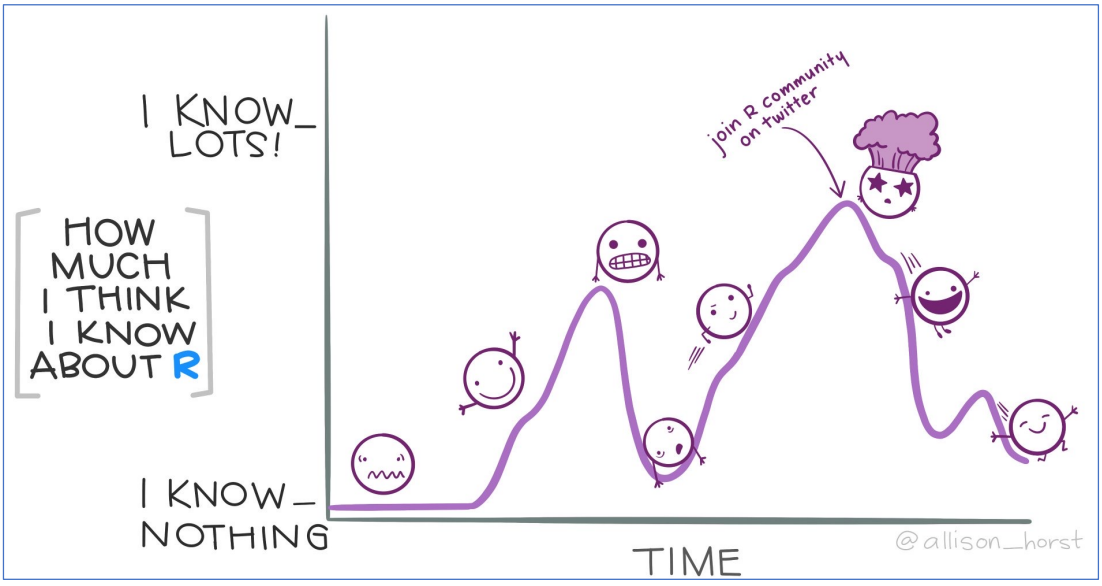
These observations performed using EDA only; no statistical models applied.

Results can be easily formulated into an actionable plan designed to reduce churn rate.

Contents

How to learn R?

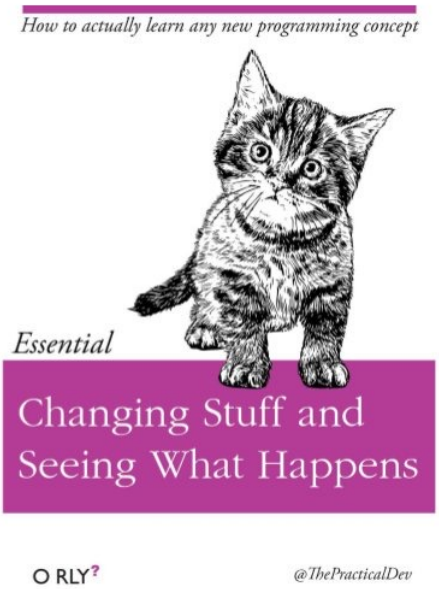
For those of you who are interested to learning a new programming language



That how I learn R and C

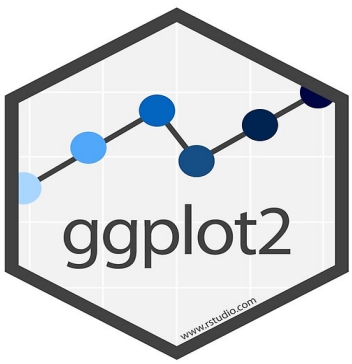
How to learn R?

Here are some important things to keep in mind as you learn (these are joke book covers):

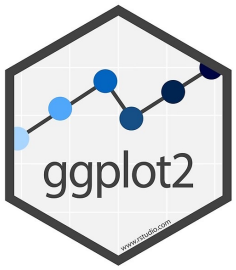


Data Visualization using ggplot2

Data visualization using ggplot2 package for the churn dataset from the lecture of Week 2.



Visualization via ggplot2



ggplot2: https://ggplot2.tidyverse.org/reference/

Plot basics

All ggplot2 plots begin with a call to `ggplot()`, supplying default data and aesthetic mappings, specified by `aes()`. You then add layers, scales, coords and facets with `+`. To save a plot to disk, use `ggsave()`.

`ggplot()`	Create a new ggplot
`aes()`	Construct aesthetic mappings
`` `+`(<gg>) `%+%` ``	Add components to a plot
`ggsave()`	Save a ggplot (or other grid object) with sensible defaults
`qplot()` `quickplot()`	Quick plot

Layers

Geoms

A layer combines data, aesthetic mapping, a geom (geometric object), a stat (statistical transformation), and a position adjustment. Typically, you will create layers using a `geom_` function, overriding the default position and stat if needed.

`geom_abline()` `geom_hline()` `geom_vline()`	Reference lines: horizontal, vertical, and diagonal
`geom_bar()` `geom_col()` `stat_count()`	Bar charts

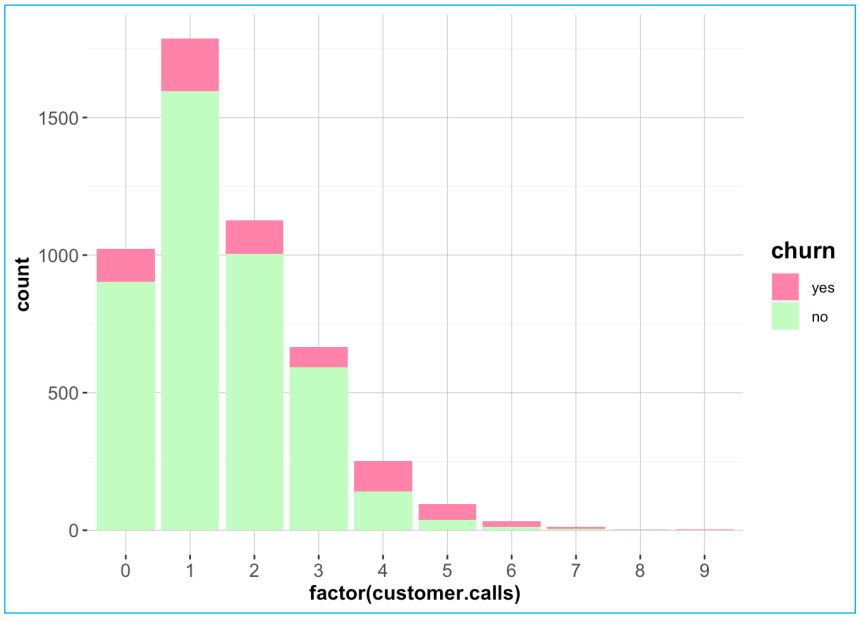
Visualization via ggplot2

Code template example:

```
ggplot( data = <DATA> ) +
    <GEOM_FUNCTION>(
         mapping = aes( <MAPPINGS> ),
         stat = <STAT>,
         position = <POSITION>
    ) +
    <COORDINATE_FUNCTION> +
    <FACET_FUNCTION>
```

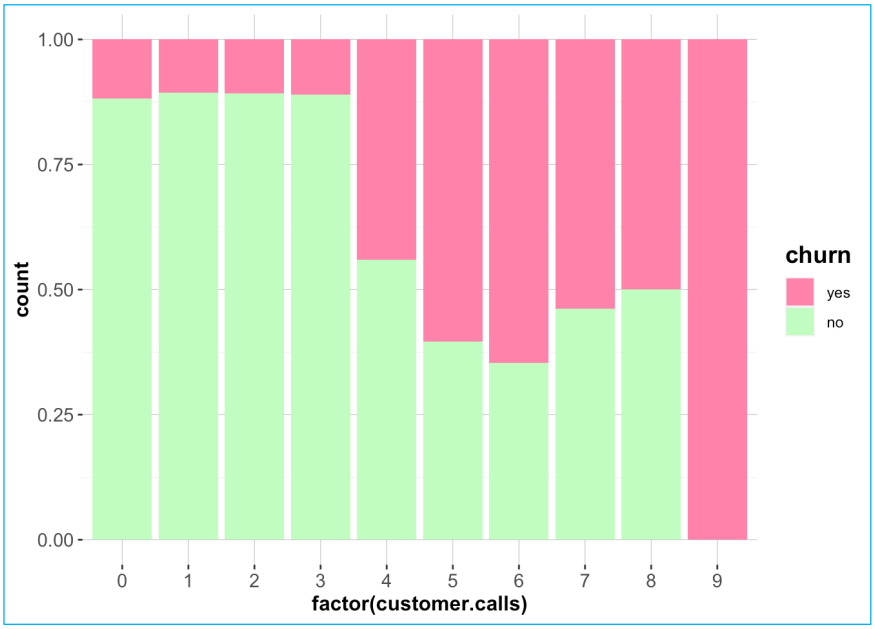
Bar Chart/Histogram

```
ggplot( data = churn ) +
    geom_bar( aes( x = customer.calls, fill = churn ), position = "stack" )
```



Proportional Bar/Histogram

```
ggplot( data = churn ) +
    geom_bar( aes( x = customer.calls, fill = churn ), position = ”fill" )
```



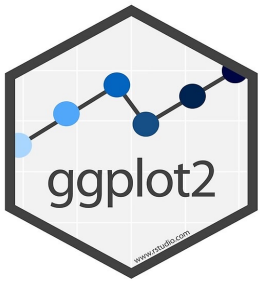
Main difference in R code?

The position adjustment:

1. position = "stack",
2. position = "fill".

Note also the fill = option as an aesthetic for coloring the bar chart.

More info



See the "Exercises_week2.Rmd" file at Canvas for more examples.

More info about **ggplot2**:

https://www.datacamp.com/courses/data-visualization-with-ggplot2-1

https://ggplot2.tidyverse.org/reference/

Cheat Sheet for ggplot2:

https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Misc/data-visualization-2.1.pdf