Data visualization:

Project: IMDB movie analysis

Purpose: Analyze and generate insight from IMDB dataset, generate a dashboard for movie audience.

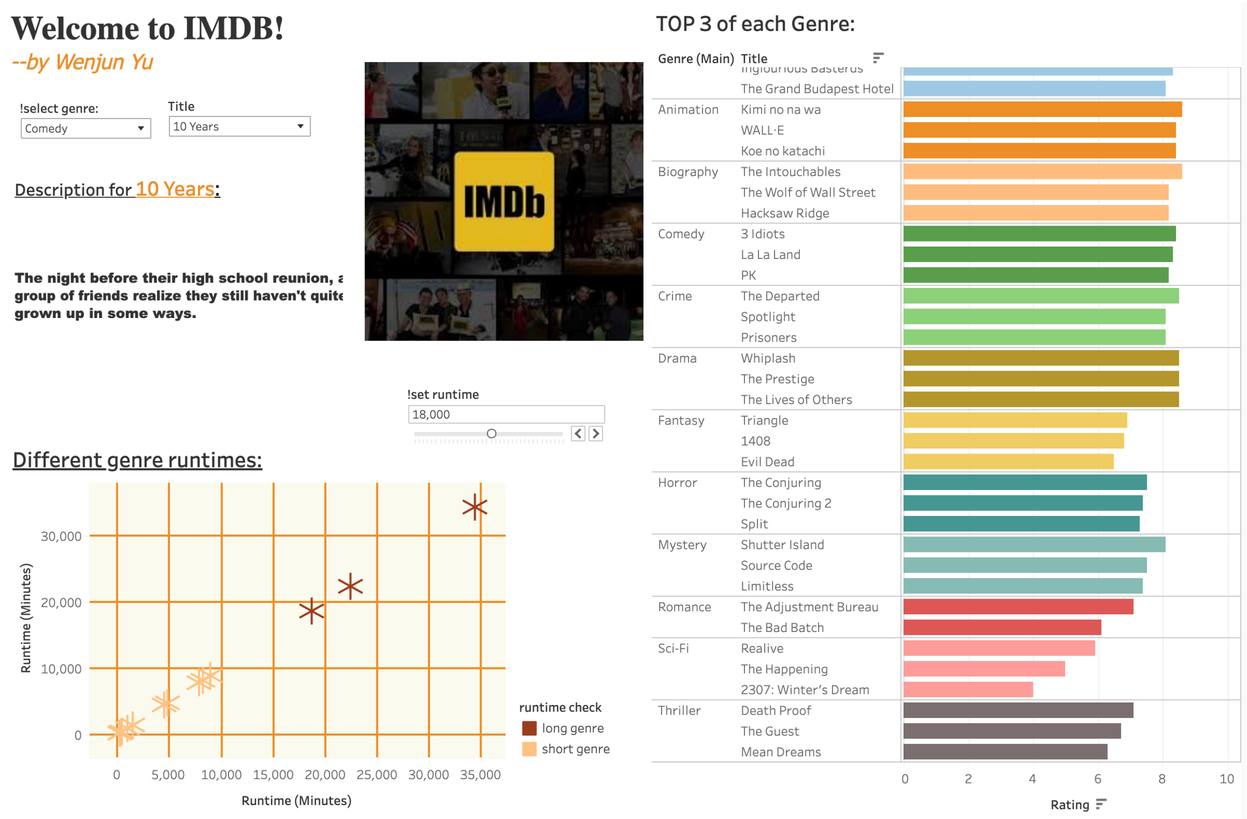
Tool: Tableau

Data source: Dataset from IMDb, which is an online database of information related to films, television programs, home videos, video games, and streaming content online.

Methodology:

1) advanced features of Tableau, including parameter, calculated fields, filter.

https://public.tableau.com/profile/wenjun.yu#!/vizhome/WenjunYu_IMDB/DashboardIMDB



Welcome to IMDB!

--by Wenjun Yu

!select genre: Comedy

Title: 10 Years

Description for 10 Years:

The night before their high school reunion, a group of friends realize they still haven't quite grown up in some ways.

!set runtime: 18,000

Different genre runtimes:

TOP 3 of each Genre:

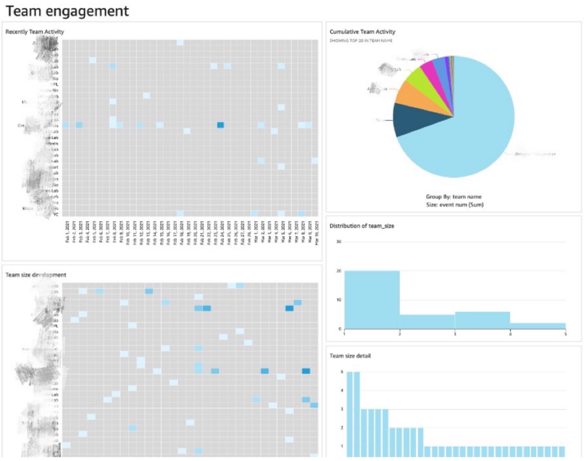
Genre (Main)	Title
	Inglourious Basterds
	The Grand Budapest Hotel
Animation	Kimi no na wa
	WALL·E
	Koe no katachi
Biography	The Intouchables
	The Wolf of Wall Street
	Hacksaw Ridge
Comedy	3 Idiots
	La La Land
	PK
Crime	The Departed
	Spotlight
	Prisoners
Drama	Whiplash
	The Prestige
	The Lives of Others
Fantasy	Triangle
	1408
	Evil Dead
Horror	The Conjuring
	The Conjuring 2
	Split
Mystery	Shutter Island
	Source Code
	Limitless
Romance	The Adjustment Bureau
	The Bad Batch
Sci-Fi	Realive
	The Happening
	2307: Winter's Dream
Thriller	Death Proof
	The Guest
	Mean Dreams

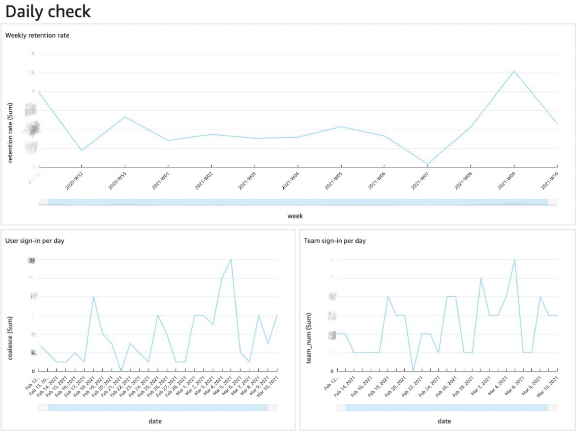
Project: Customer Success Project

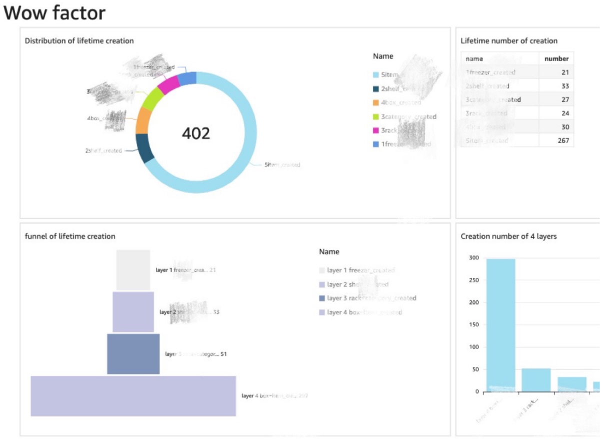
Purpose: Measure customer behavior to improve system design and reduce potential customer churn.

Tool: AWS QuickSight

Data Source: user engagement data collected from company platform.
Methodology:
1) Customized complex SQL Query, including multiple table joining, group by, order by, window function, CTE, multiple level of subquery, Case When, aggregate functions, statistical functions, date related functions, calculation related functions
2) multiple types of chart, such as Line chart, Bar chart, Pie chart, distribution chart, waterfall chart, funnel chart, donut chart and heatmap

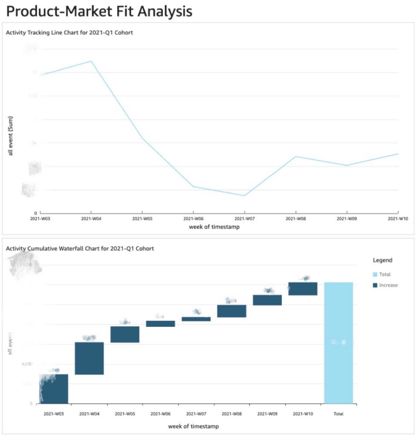





Wow factor
Distribution of lifetime creation
402
Name
funnel of lifetime creation
Name
Creation number of 4 layers
300
250
200
150
100
50
0

Lifetime number of creation

name	number
1freezer_created	21
2shelf_created	33
3category_created	27
3rack_created	24
4box_created	30
5item_created	267



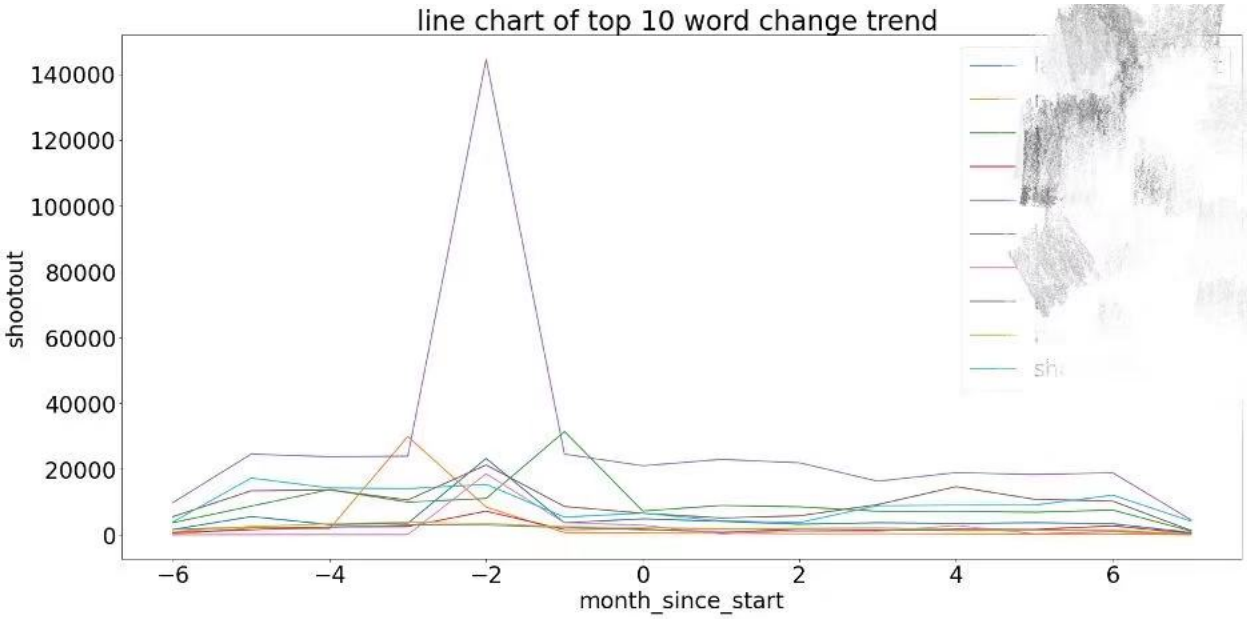
Big Data:

Project: Twitter Hadoop Project

Purpose: Testing for chilling effects in online social media platforms.

Tool: Python, Hadoop, MapReduce, Linux
Data source: Our unique dataset, the Twitter Decahose, which contains a random 10% sample of all Tweets, comes from a privileged agreement with a private North American University. And there is research paper to show how such a Decahose is a truly representative sample of the

Firehose (overall Twitter data), and how this data can be used in any Twitter-related analysis. It contains total one-year tweets of Twitter. Also, we have a special word list which NSA admittedly monitors on social media.
Methodology:
1) NLP preprocess tweets
2) Use Python MapReduce to build ETL data pipeline for this 80 TB dataset
3) Generate insight about word usage trend, top words for different regions



Data Science:

Project: Abalone Dataset Analysis

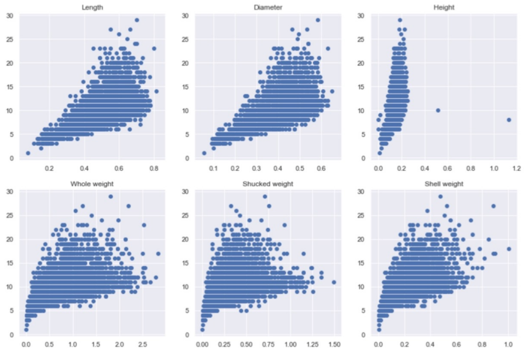
Purpose: Analyze abalone dataset, come up with some conclusions from general analysis, then use measurements and pvalues to support.

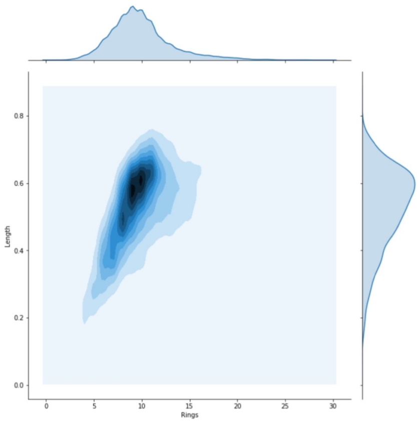
Tool: Python, Pandas, Matplotlib
Data Source: Data comes from an original (non-machine-learning) study: Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)

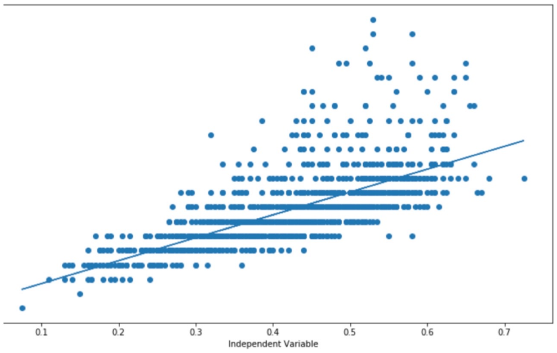
Methodology:
1) Read in and clean dataset: handling null values, encoding categorical variables (two methods: get_dummies and map),binning.

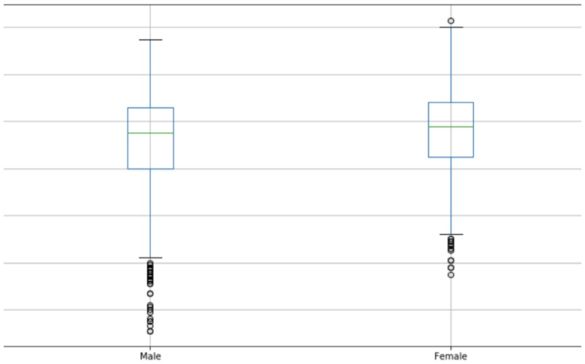
2) Exploratory data analysis: Examing the distribution of features, Feature normalization and scaling, Examining and plotting correlation between variables, Detecting and Handling outliers
3) Draw conclusions based on exploratory data analysis. Each of conclusion is supported with visualizations and pvalues and any other appropriate measurments for dataset.

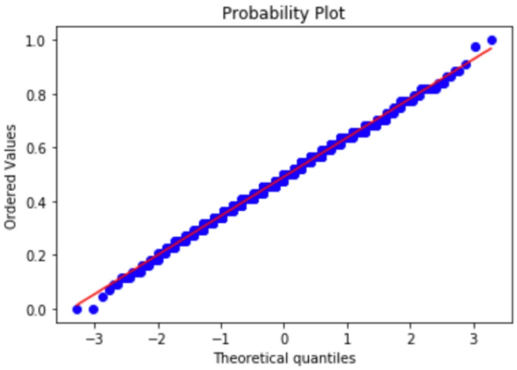
- boxplot, Anova test, T test
- Mann Whitney U-test, median_test, hist plot
- seaborn, Pearsonr correlation coefficient
- Multiple linear Regression
- QQ plot, normaltest

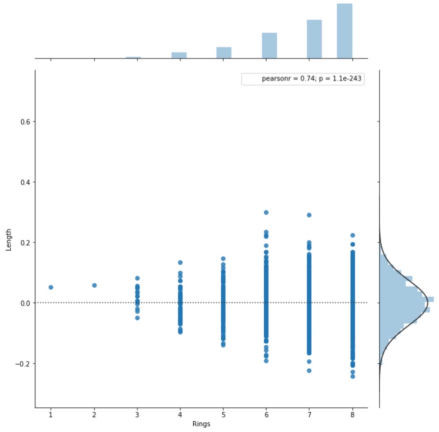

Length
Diameter
Height
Whole weight
Shucked weight
Shell weight
0
5
10
15
20
25
30
0.2
0.4
0.6
0.8
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
0.00
0.25
0.50
0.75
1.00
1.25
1.50
0.0
0.2
0.4
0.6
0.8
1.0


Length
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
30
Rings


0.1
0.2
0.3
0.4
0.5
0.6
0.7
Independent Variable


Male
Female


Probability Plot
Ordered Values
0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3
Theoretical quantiles


pearsonr = 0.74; p = 1.1e-243
Length
−0.2
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
8
Rings

Project: Conversion rate Analysis

Purpose: Predict conversion rate Come up with recommendations for the product team and the marketing team to improve conversion rate.

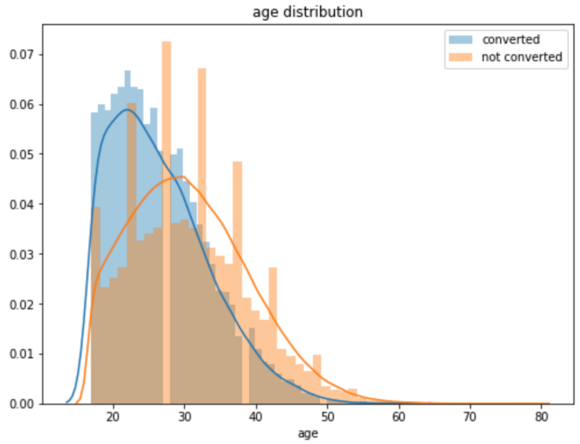
Tool: Python, Pandas, Matplotlib, Sklearn

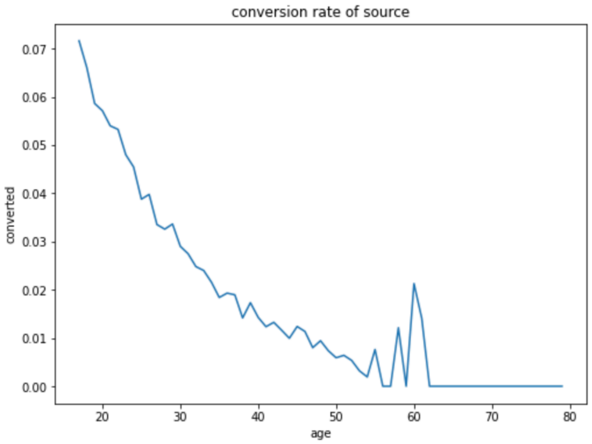
Data Source: A collection of Data Science Take Home Challenge

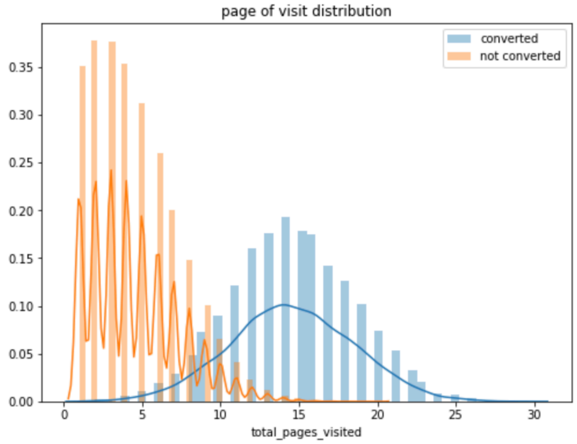
Methodology:

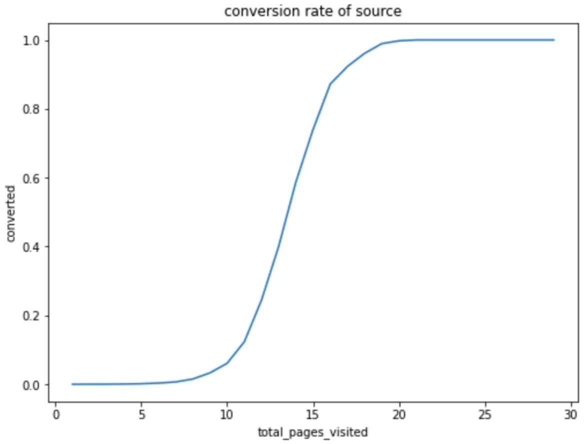
1) machine learning models:

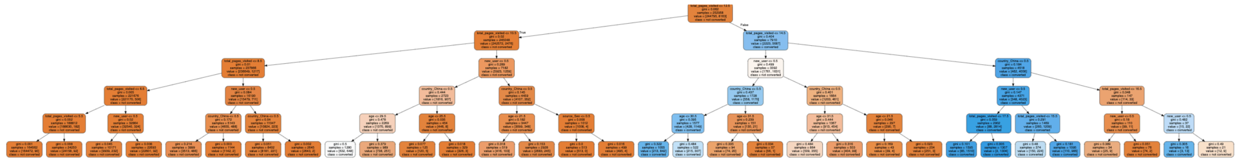
- Decision Tree (gini,max_depth,min_impurity_decrease,GridSearchCV)
- Random Forest (AUC, roc_curve, confusion matrix，h2o frame)

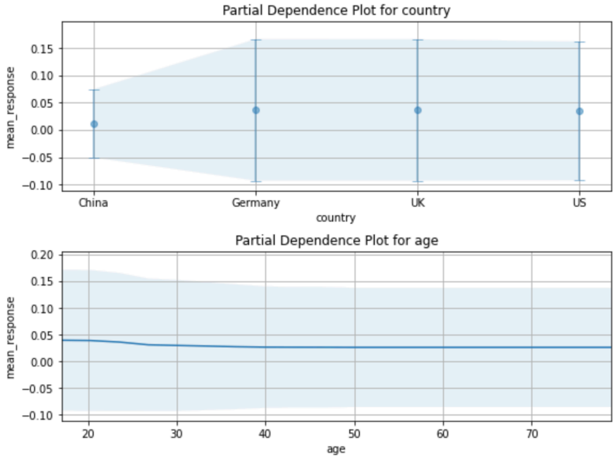


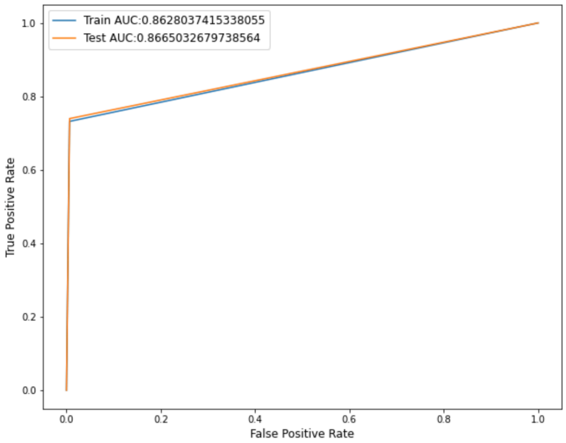









Partial Dependence Plot for country
mean_response
0.15
0.10
0.05
0.00
−0.05
−0.10
China
Germany
UK
US
country
Partial Dependence Plot for age
mean_response
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
20
30
40
50
60
70
age


Train AUC:0.8628037415338055
Test AUC:0.8665032679738564
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate