Breaking Down Data Science Life-Cycle.

IBM Developer

Breaking Down Data Science Life Cycle.

Episode #1

23-Oct

- Discovery & Gathering of Unstructured Data.
- Data Understanding.

Episode #2

21-Nov

- Data Preparation.
- ICP for Data.

Episode #3

26-Dec

- Machine Learning on Data.
- Modeling and Deployment.

Data Discovery & Gathering

Mashael AlMuhanna
Data Governance Specialist

IBM **Developer**

What is Data ?

How is it Generated ?

Data refers to information that is machine-readable as opposed to human-readable.

The importance of Data

Data driven HR.

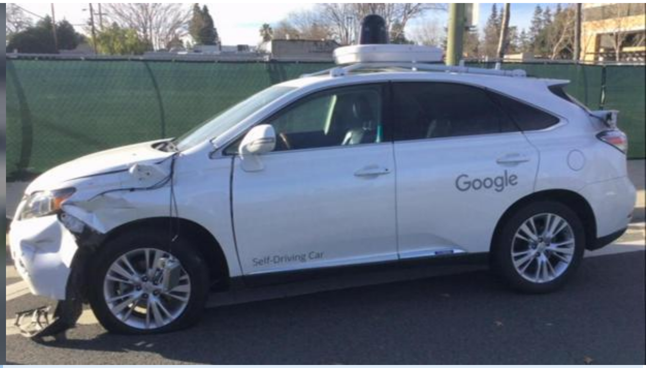
Companies target

Data driven Strategies

Relate to customers

Data is Everywhere

Influencing What We Do



Netflix provides personalized recommendations

Waze provides a personalized driving experience

Uber delivers food that you like and is the right temperature

Self driving cars react to changing conditions

ALL based on DATA

Data-driven cultures realize higher business returns

Achieve Differentiation

Manufacturing
Predictive maintenance,
production output & inventory

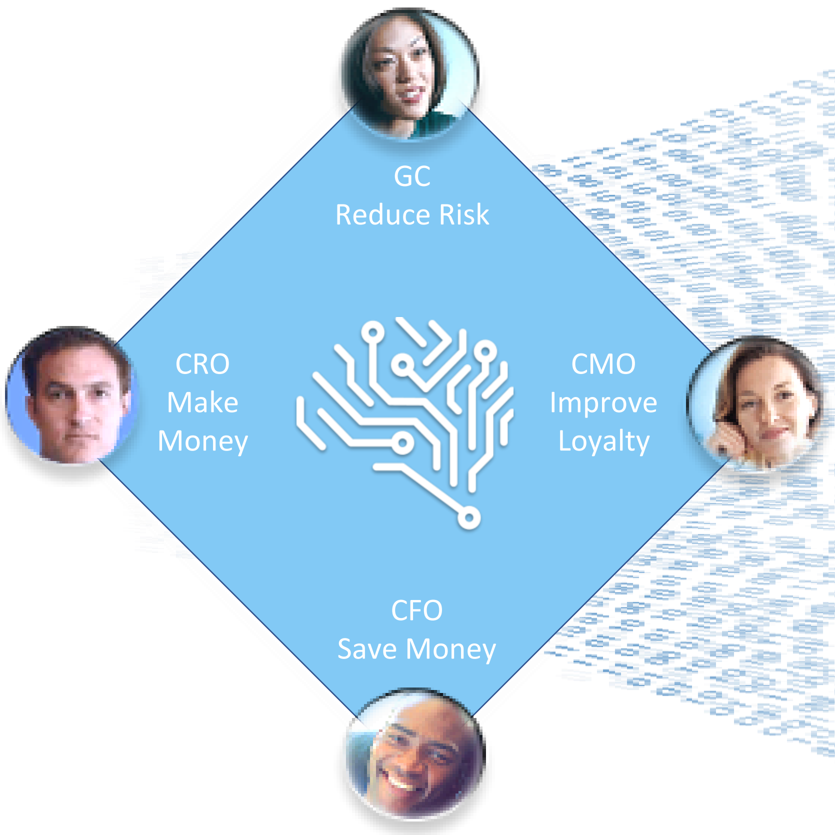
Banking & Financial
Reveal trading behavior,
regulatory compliance

Retail
Dynamic pricing and
predictive merchandising

Healthcare
Accuracy of diagnosis
and regulatory compliance

Telecom & Media
Predictive customer
Experience and loyalty

Drive 6% Greater Productivity*



GC
Reduce Risk

CRO
Make
Money

CMO
Improve
Loyalty

CFO
Save Money

Apply Technology 13:1* ROI

Cloud
data agility
and efficiencies

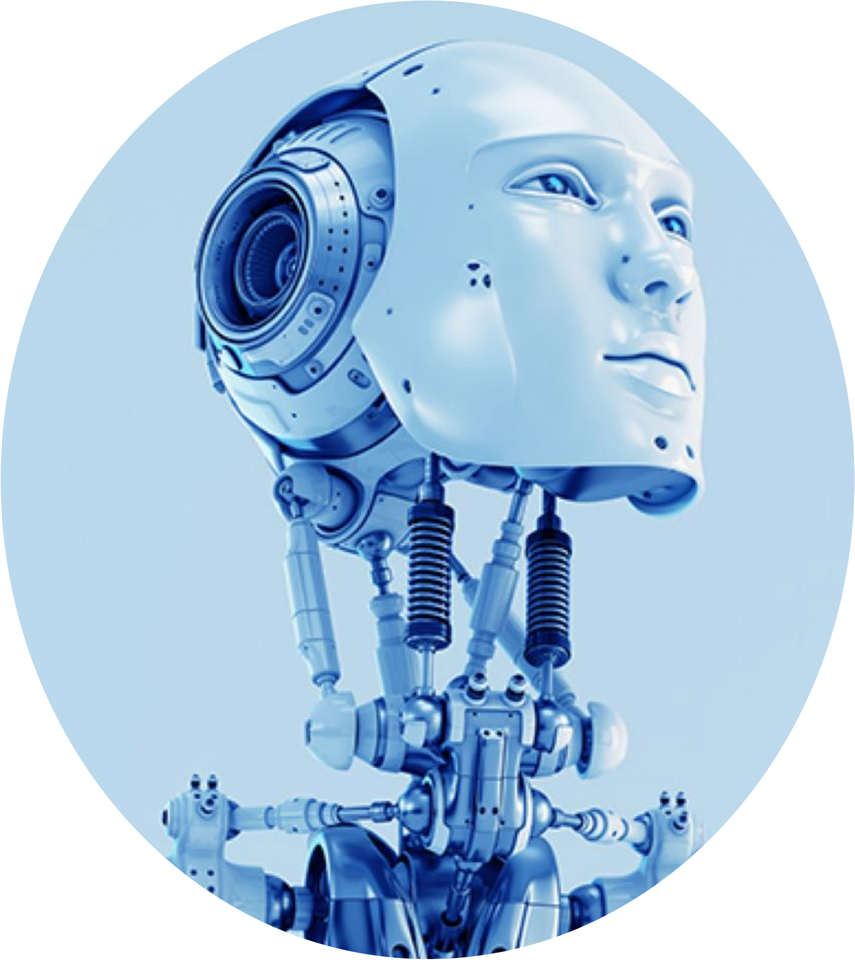
Mobile-IoT
real-time and
flexible data access

Open Source
speed innovation
and data exploitation

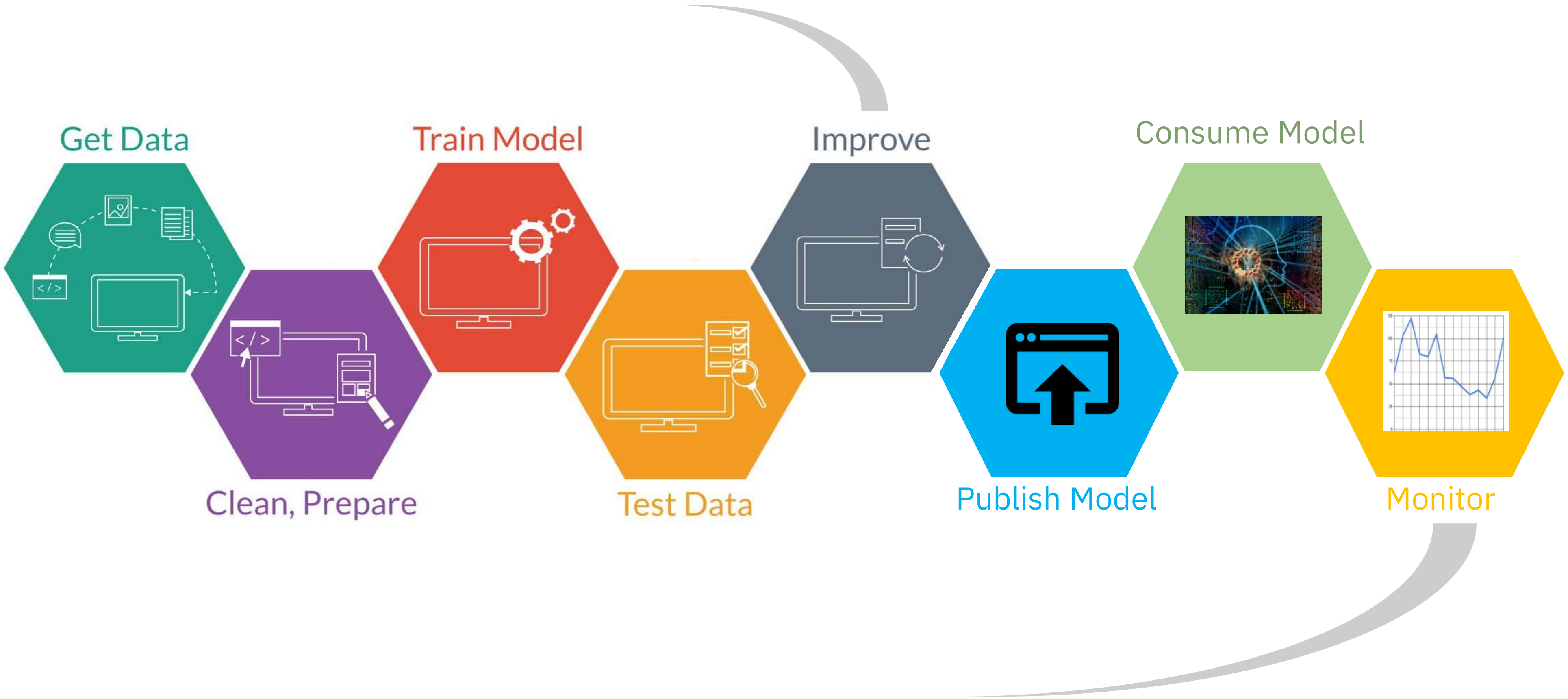
Artificial Intelligence
scale discovery
of hidden insights

* MIT Slone School of Management Study of 330 firms

Data-related challenges are hindering 96% of organizations from achieving AI

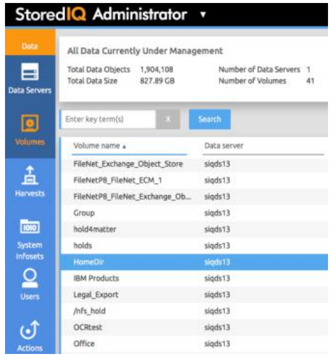


What Real-World AI Actually Looks Like



We make data simple
and accessible

Discover & Search for Data, Refine & Act



All Data Currently Under Management

Total Data Objects	1,904,108	Number of Data Servers	1
Total Data Size	827.89 GB	Number of Volumes	41

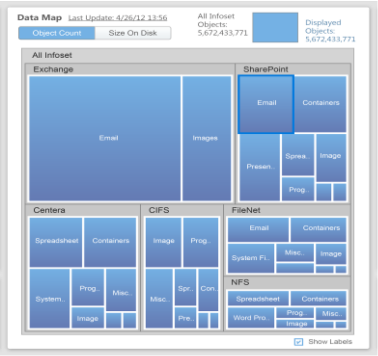
Volume name ▲	Data server
FileNet_Exchange_Object_Store	siqds13
FileNetP8_FileNet_ECM_1	siqds13
FileNetP8_FileNet_Exchange_Ob...	siqds13
Group	siqds13
hold4matter	siqds13
holds	siqds13
HomeDir	siqds13
IBM Products	siqds13
Legal_Export	siqds13
/nfs_hold	siqds13
OCRtest	siqds13
Office	siqds13

Select Data Sources

Add source(s)

Index data in place

Infosets: group all volumes you want to search in



Discover Your Data

Visualizations show:
- Types of data stored
- Allocation by Date
- Allocation by Size

Discover where your oldest, biggest or least used data resides

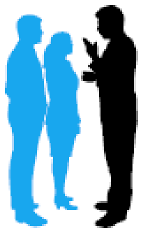
Utilize overlays to highlight potential sensitive data

Search & Refine

Identify specific data based in your requirements and search criteria

Create own Filters based on Company Policies and Common sense to find datasets to act on

Create result data sets



Report Findings

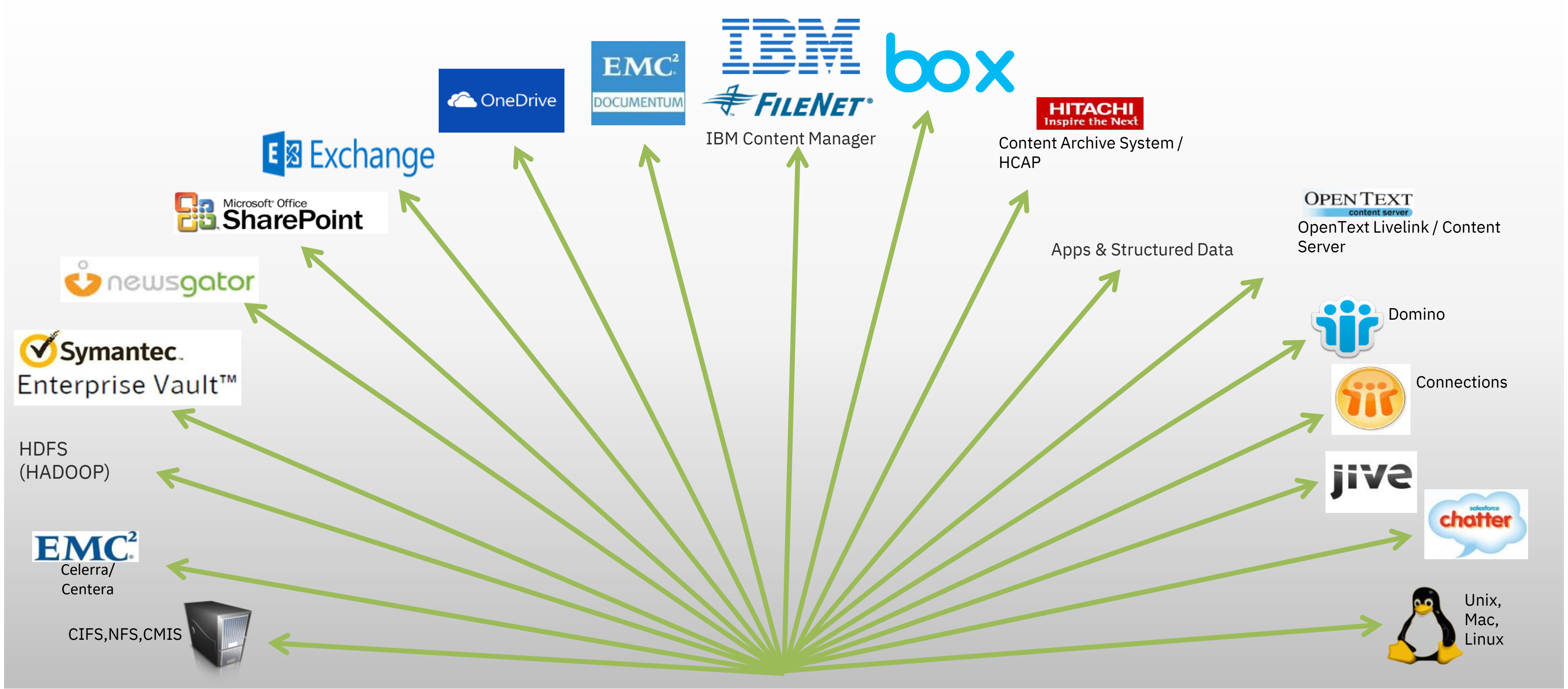
Reports of result data sets

Notify data owner

Act on Result Sets

Manage in place
- Delete
- Move
- Export
- Retain

Open Architecture



IBM Content Manager

Content Archive System / HCAP

Apps & Structured Data

OpenText Livelink / Content Server

Domino

Connections

HDFS (HADOOP)

Celerra/ Centera

CIFS,NFS,CMIS

Unix, Mac, Linux

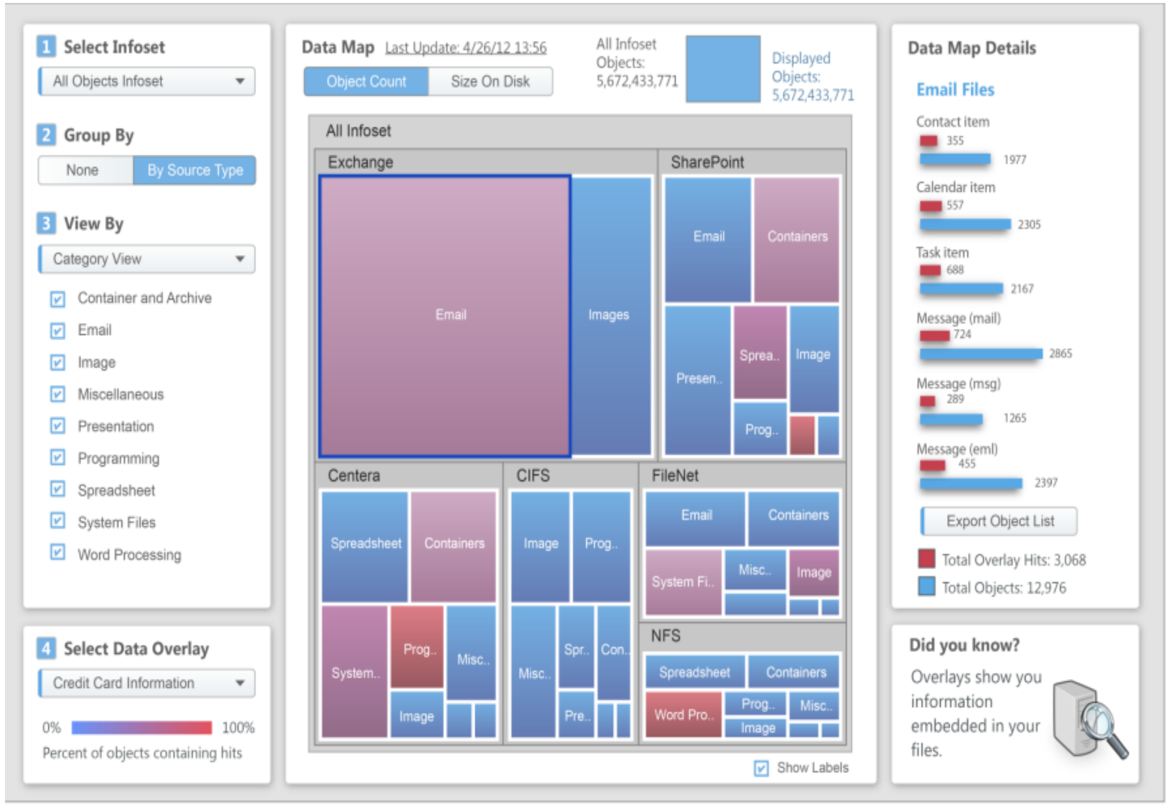
Support for 100+ data sources and 450+ file types

Discover, Analyze & Act to Govern Information

- Discover
  - Unstructured data across your enterprise

- Analyze
  - Identify sensitive & critical information:
    - National ID
    - Passport numbers
    - Name
    - Addresses
    - Etc.
  - Advanced search capabilities to help find country specific data

- Act
  - Declare as records, delete, move…



1 Select Infoset

All Objects Infoset ▾

2 Group By

None | By Source Type

3 View By

Category View ▾

☑ Container and Archive
☑ Email
☑ Image
☑ Miscellaneous
☑ Presentation
☑ Programming
☑ Spreadsheet
☑ System Files
☑ Word Processing

4 Select Data Overlay

Credit Card Information ▾

0% 100%

Percent of objects containing hits

Data Map Last Update: 4/26/12 13:56

Object Count | Size On Disk

All Infoset Objects: 5,672,433,771

Displayed Objects: 5,672,433,771

☑ Show Labels

Data Map Details

Email Files

Contact item	355 1977
Calendar item	557 2305
Task item	688 2167
Message (mail)	724 2865
Message (msg)	289 1265
Message (eml)	455 2397

Export Object List

🟥 Total Overlay Hits: 3,068

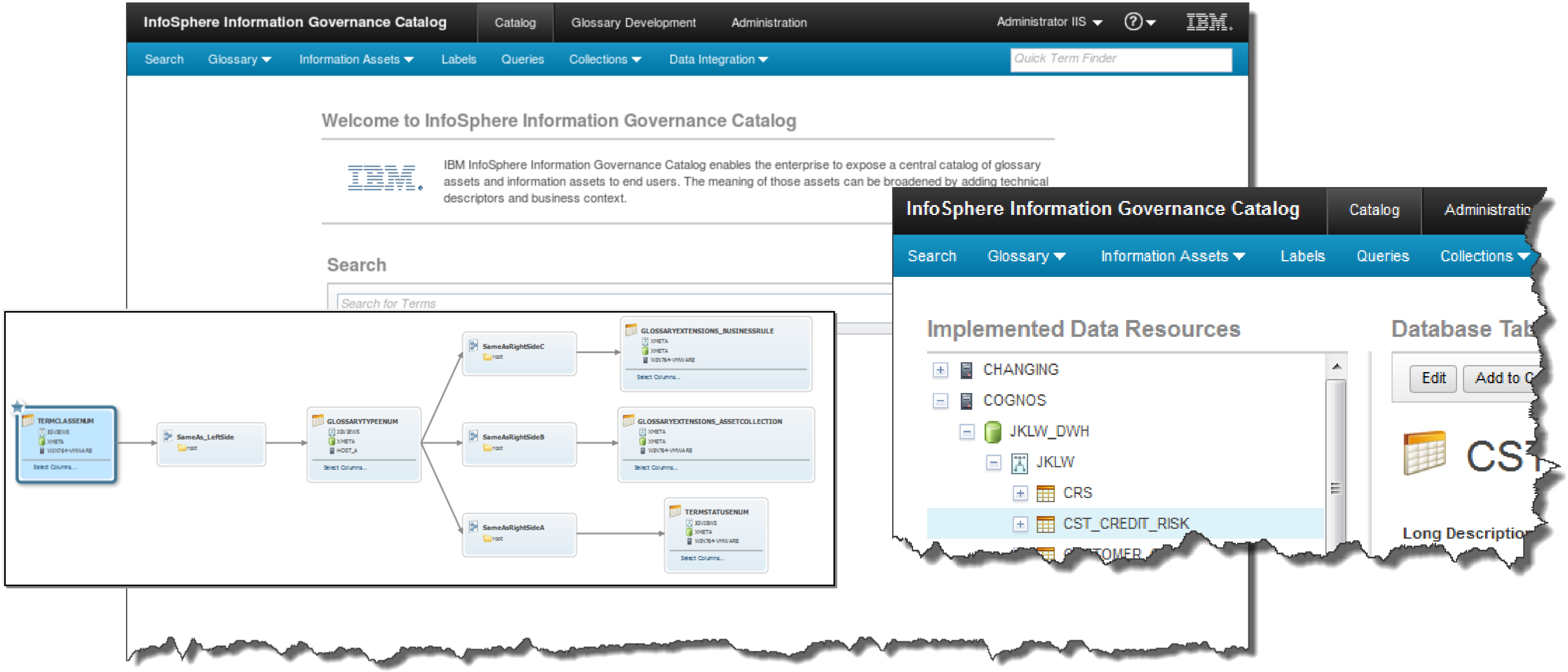
🟦 Total Objects: 12,976

Did you know?

Overlays show you information embedded in your files.

Governance Catalog

- Single interface for all governance activities
  - Establishes and promote common understanding for ALL enterprise users
  - Provides asset management and data lineage functionality

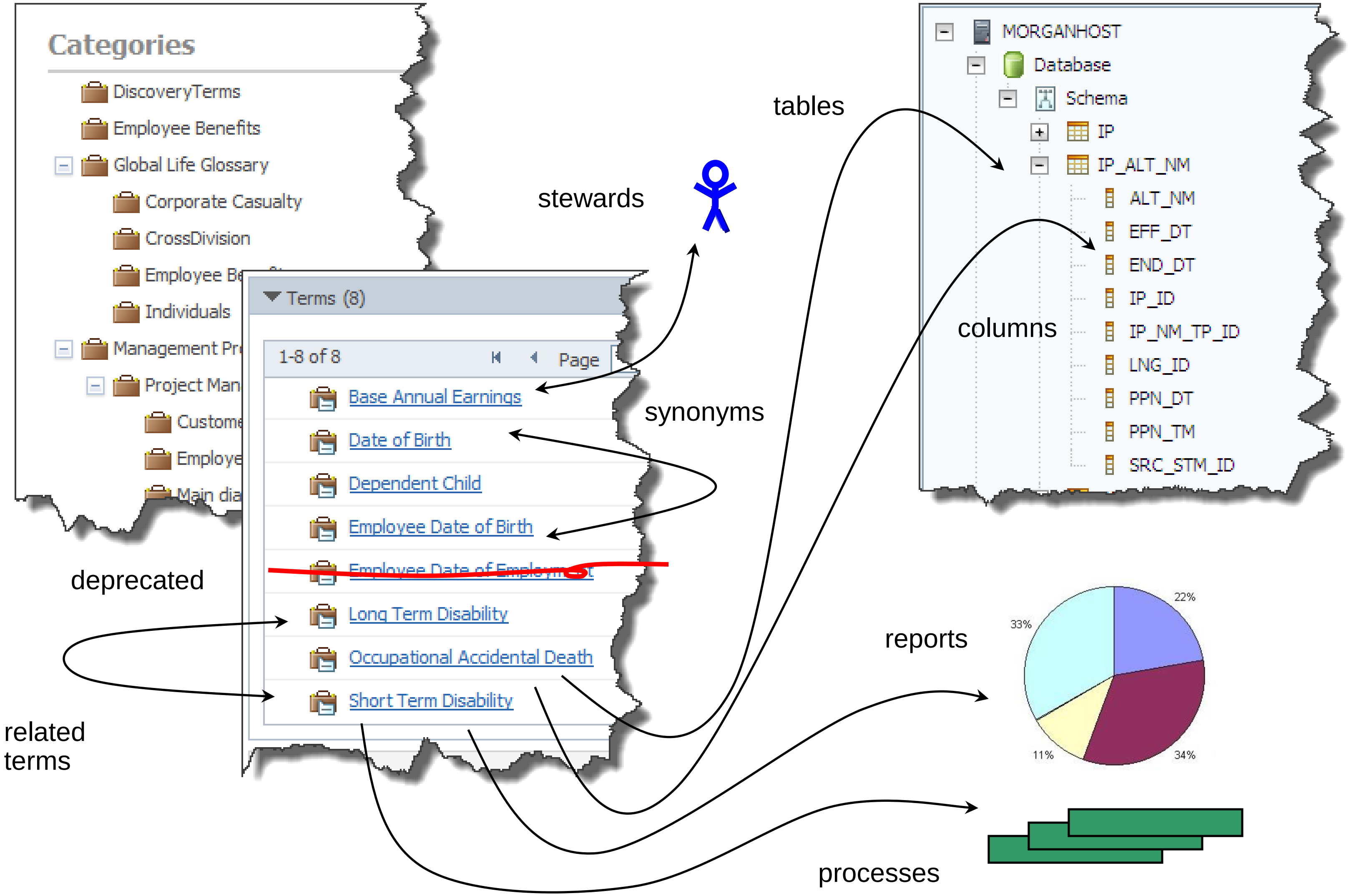


Welcome to InfoSphere Information Governance Catalog

IBM InfoSphere Information Governance Catalog enables the enterprise to expose a central catalog of glossary assets and information assets to end users. The meaning of those assets can be broadened by adding technical descriptors and business context.

Business Terms

Assets



Categories

- DiscoveryTerms
- Employee Benefits
- Global Life Glossary
  - Corporate Casualty
  - CrossDivision
  - Employee Be
  - Individuals
- Management Pr
  - Project Man
    - Custome
    - Employe
    - Main dia

Terms (8)

1-8 of 8 Page

- Base Annual Earnings
- Date of Birth
- Dependent Child
- Employee Date of Birth
- ~~Employee Date of Employment~~
- Long Term Disability
- Occupational Accidental Death
- Short Term Disability

- MORGANHOST
  - Database
    - Schema
      - IP
      - IP_ALT_NM
        - ALT_NM
        - EFF_DT
        - END_DT
        - IP_ID
        - IP_NM_TP_ID
        - LNG_ID
        - PPN_DT
        - PPN_TM
        - SRC_STM_ID

stewards

synonyms

deprecated

related terms

tables

columns

reports

processes

Data Understanding

Hissah AlMuneef
Developer Advocate

IBM **Developer**

Data exploration, data cleaning, feature engineering, pre-processing, etc…

80

20

Model building

Data Understanding

Task #1

Data Description

Task #2

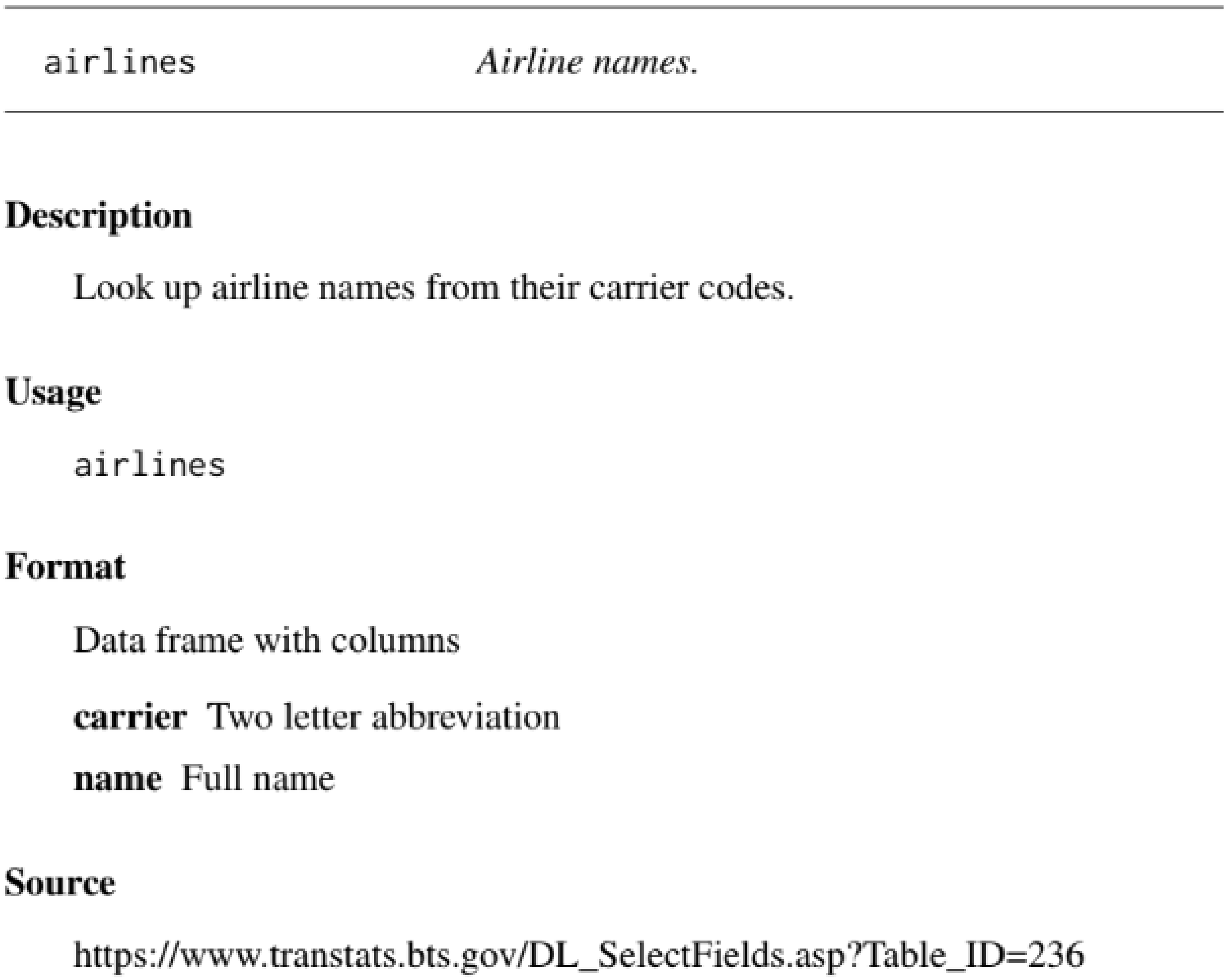
Data Exploration

Data Description

Date Description contain:

- Source of Data
- The Number of Cases
- The Number of Fields
- Description of Fields

Example for SaudiViz…



`airlines`	*Airline names.*

Description

Look up airline names from their carrier codes.

Usage

```
airlines
```

Format

Data frame with columns

carrier Two letter abbreviation

name Full name

Source

https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

Data Exploration

IBM Developer

Date Exploration Steps

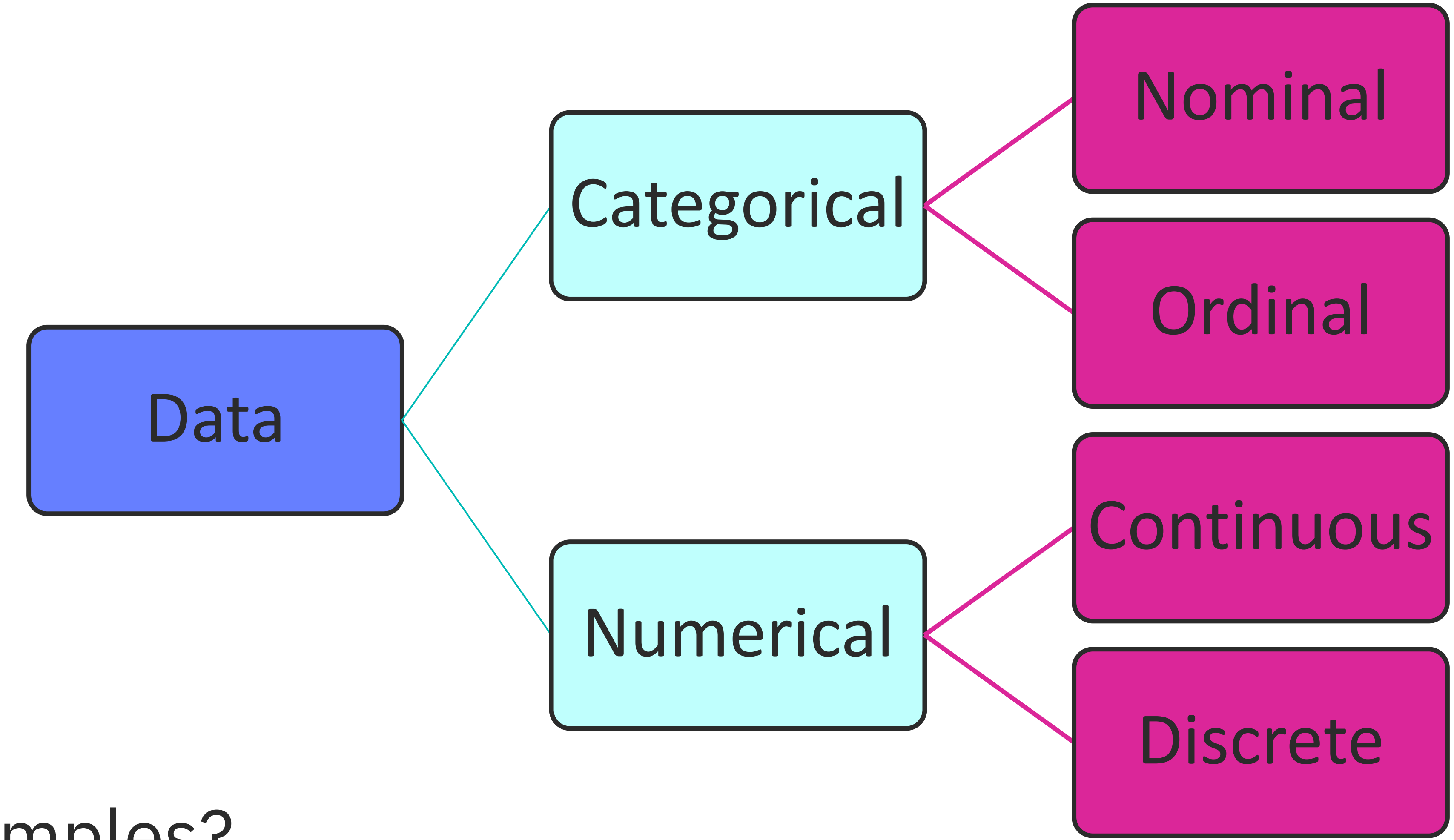
1. Identify Data Types
2. Explore Each Variable
3. Find Correlation

Step 1

Identify data types

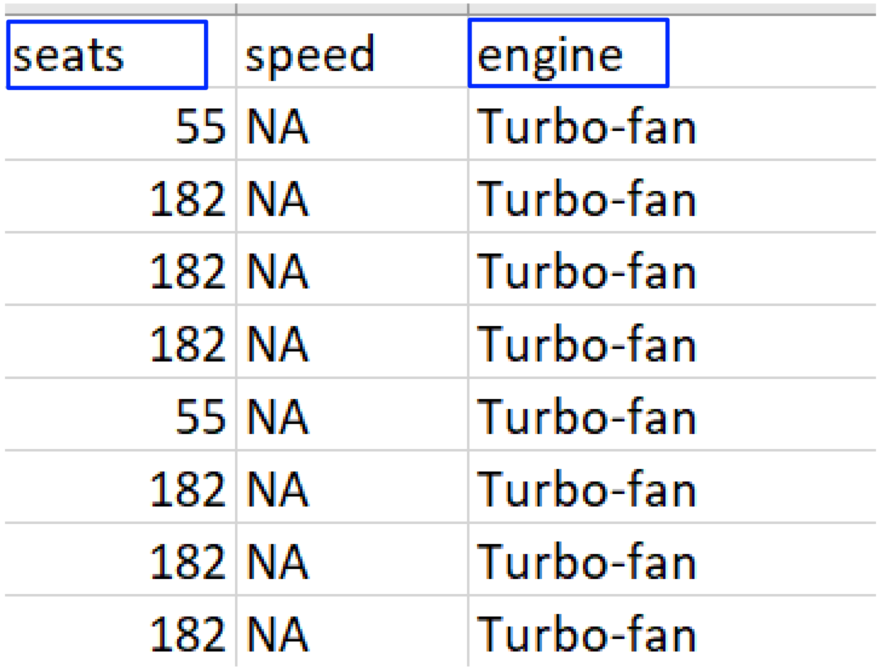
IBM **Developer**

Data Types



Other examples?

Identify data types



Numerical

Categorical

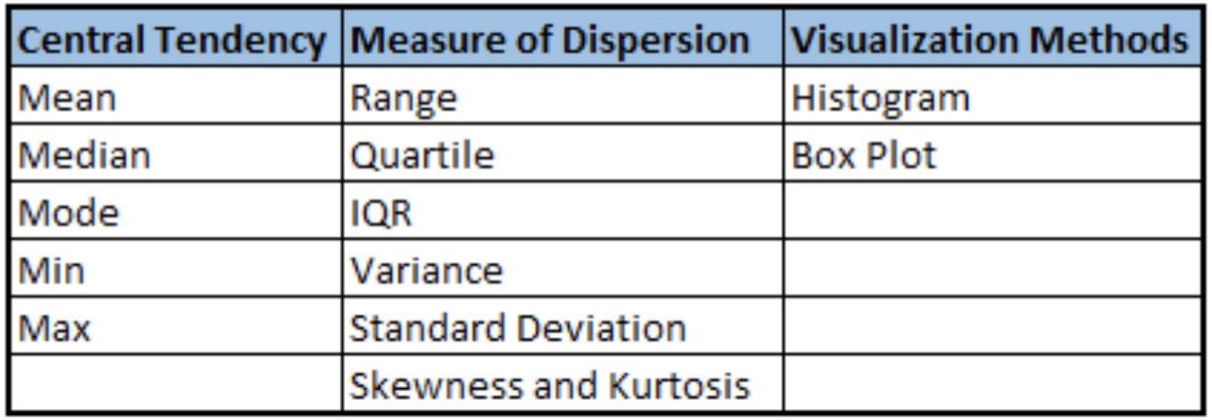
seats	speed	engine
55	NA	Turbo-fan
182	NA	Turbo-fan
182	NA	Turbo-fan
182	NA	Turbo-fan
55	NA	Turbo-fan
182	NA	Turbo-fan
182	NA	Turbo-fan
182	NA	Turbo-fan

Step 2

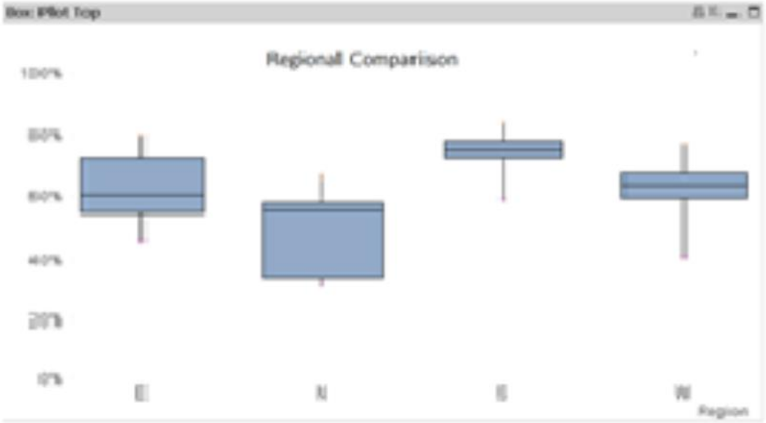
Explore Variables One by One

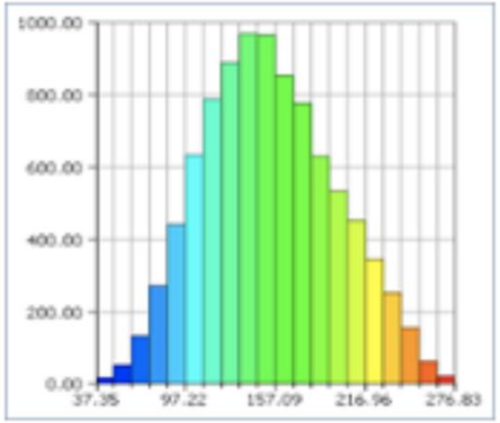
IBM **Developer**

Numerical Variable



Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	





A. Central Tendency

- Mode: most frequent measure
- Median: mid-point of an array of measures
- Mean: arithmetic average (Sum/N)

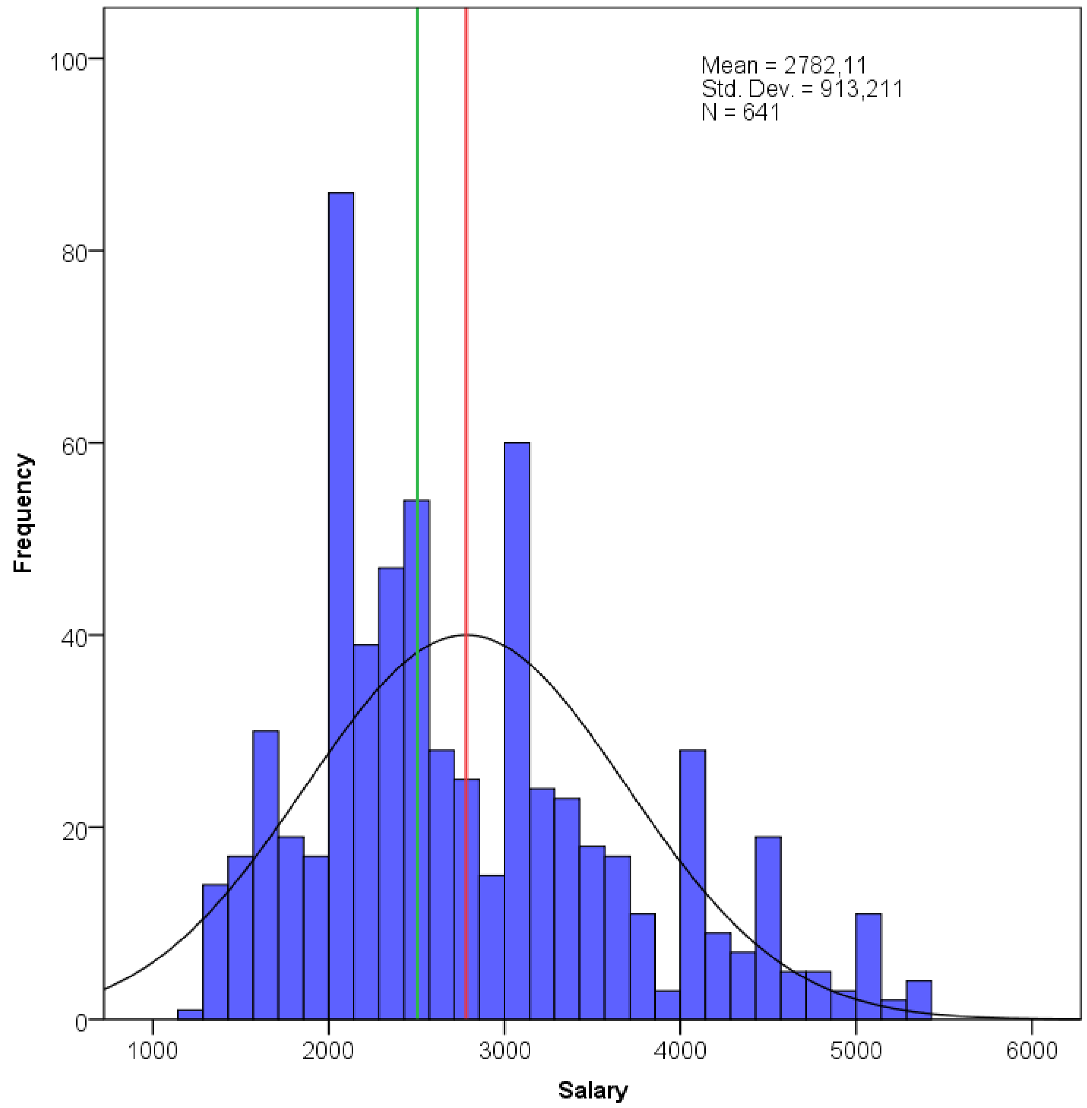
B. Variance and Standard Deviation

Standard deviation is the square root of the Variance

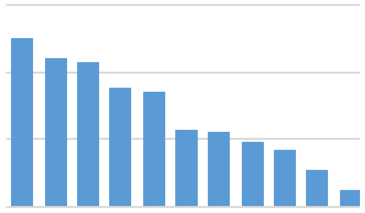
$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

x1	1	1	1
x2	1	2	2
x3	1	2	3
std	0	0.57735027	1

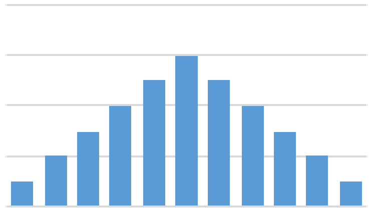
C. Histogram & Boxplot



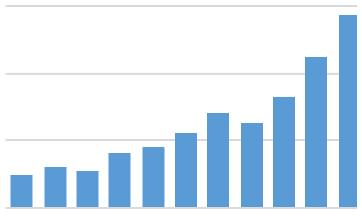
Postive Skew



Normal Distribution

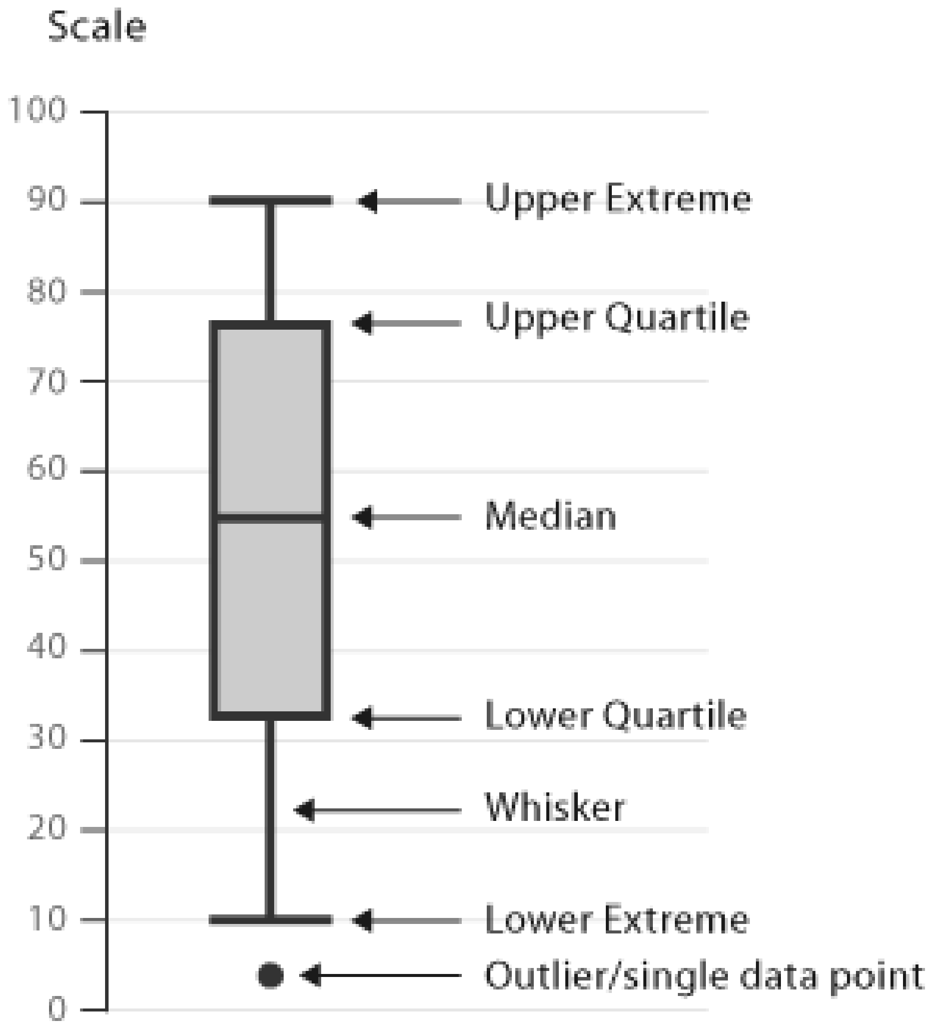


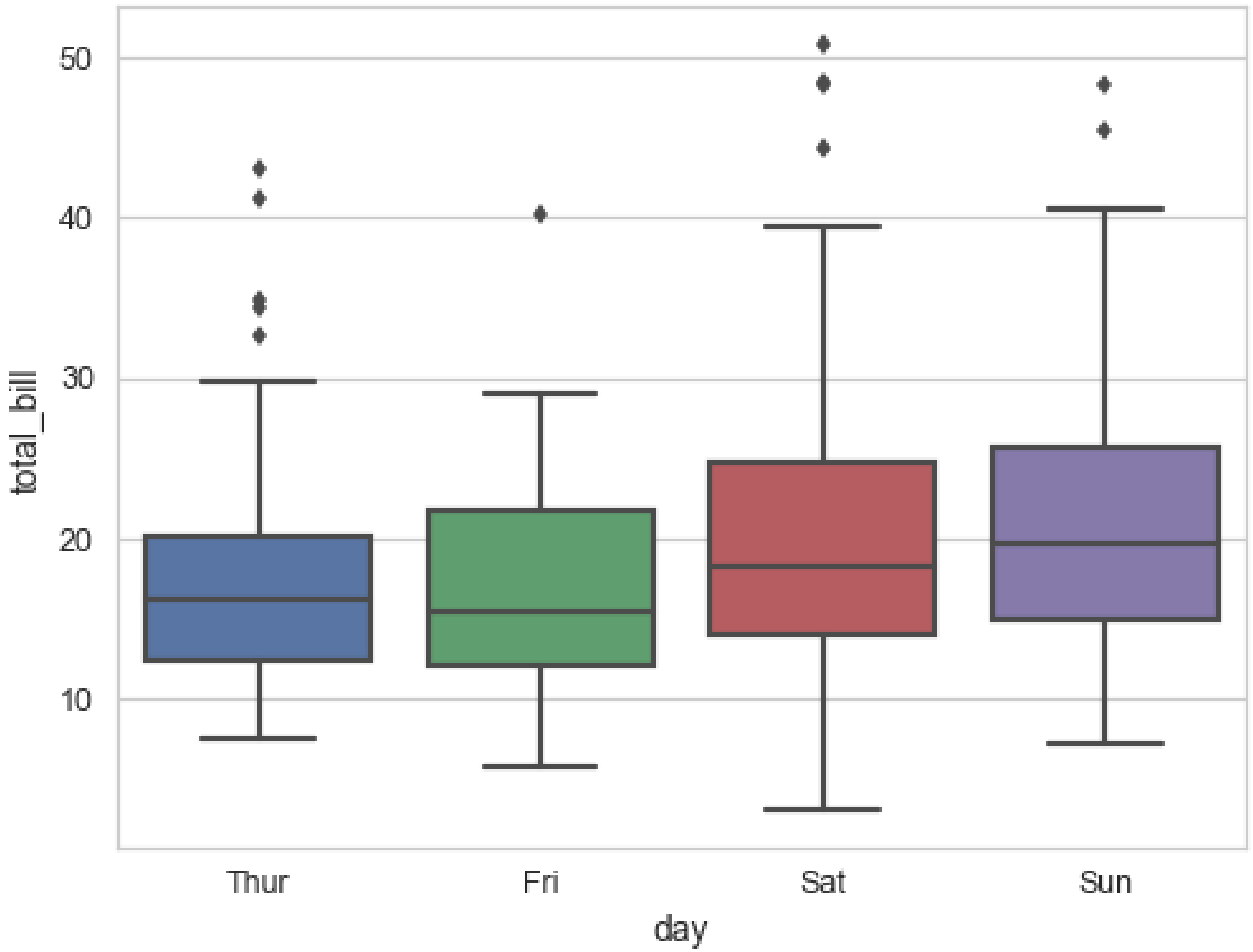
Negative Skew



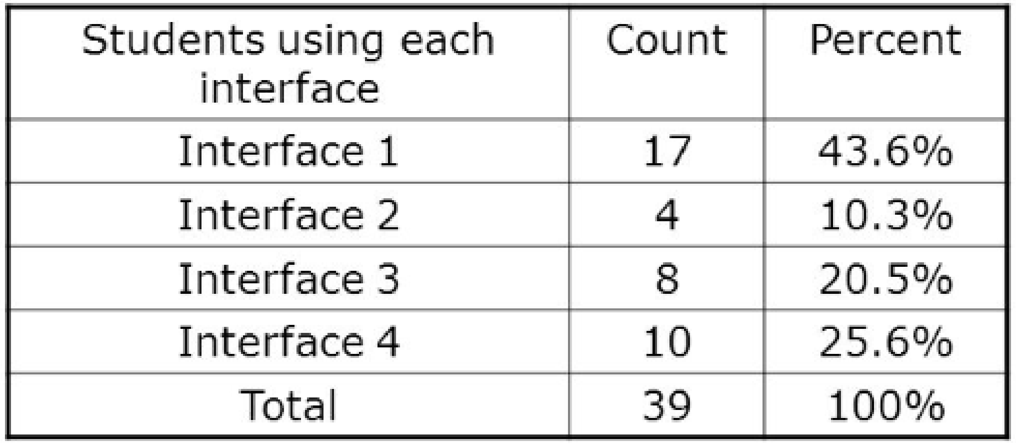
IBM **Developer**

C. Histogram & Boxplot

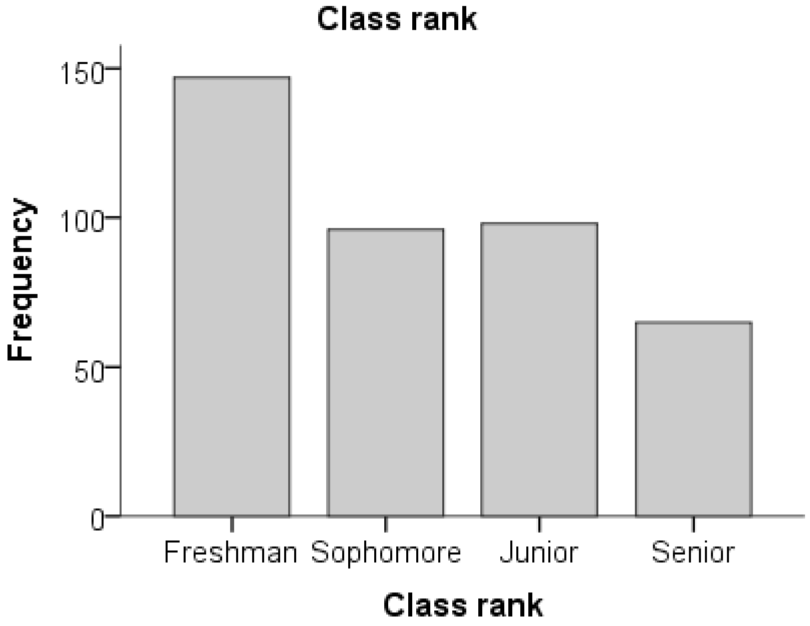




Categorical Variable

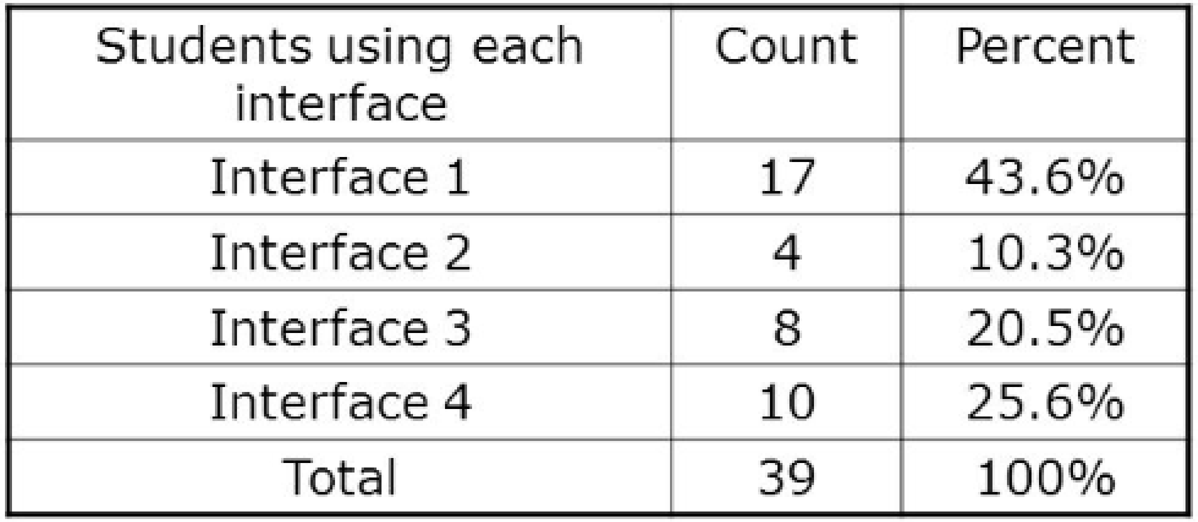


Students using each interface	Count	Percent
Interface 1	17	43.6%
Interface 2	4	10.3%
Interface 3	8	20.5%
Interface 4	10	25.6%
Total	39	100%



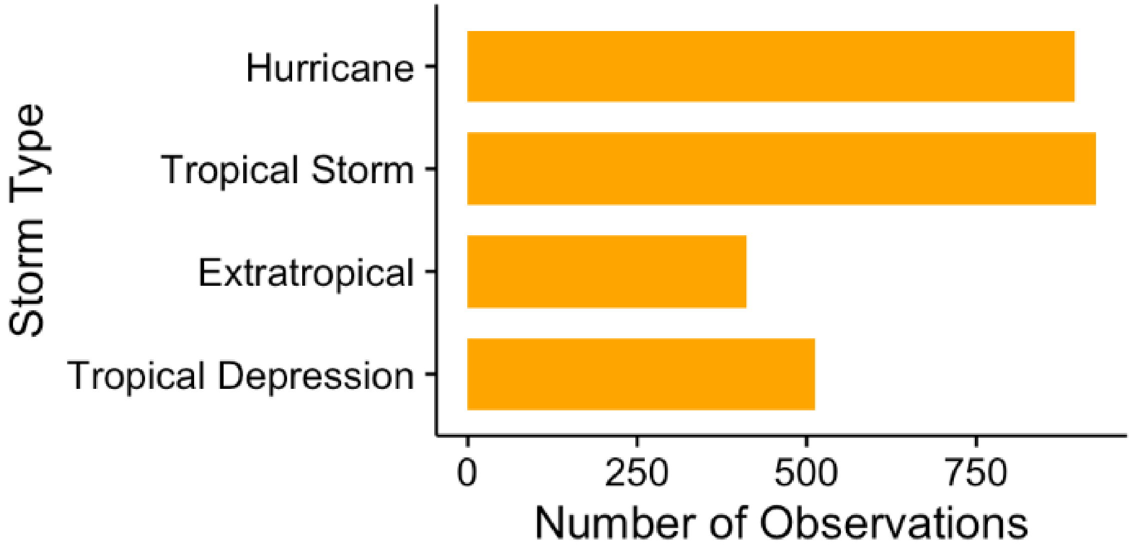
A. Numerical Summaries

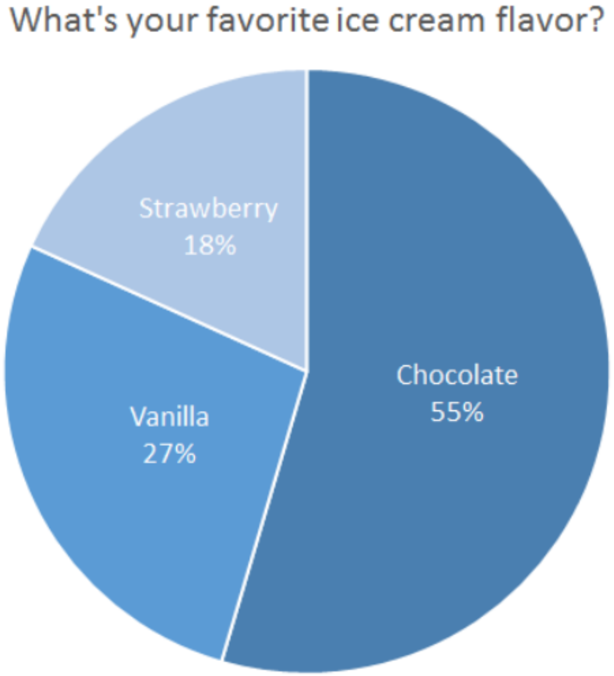
- Count
- Count%
- Mode



Students using each interface	Count	Percent
Interface 1	17	43.6%
Interface 2	4	10.3%
Interface 3	8	20.5%
Interface 4	10	25.6%
Total	39	100%

B. Bar plot & Pie plot



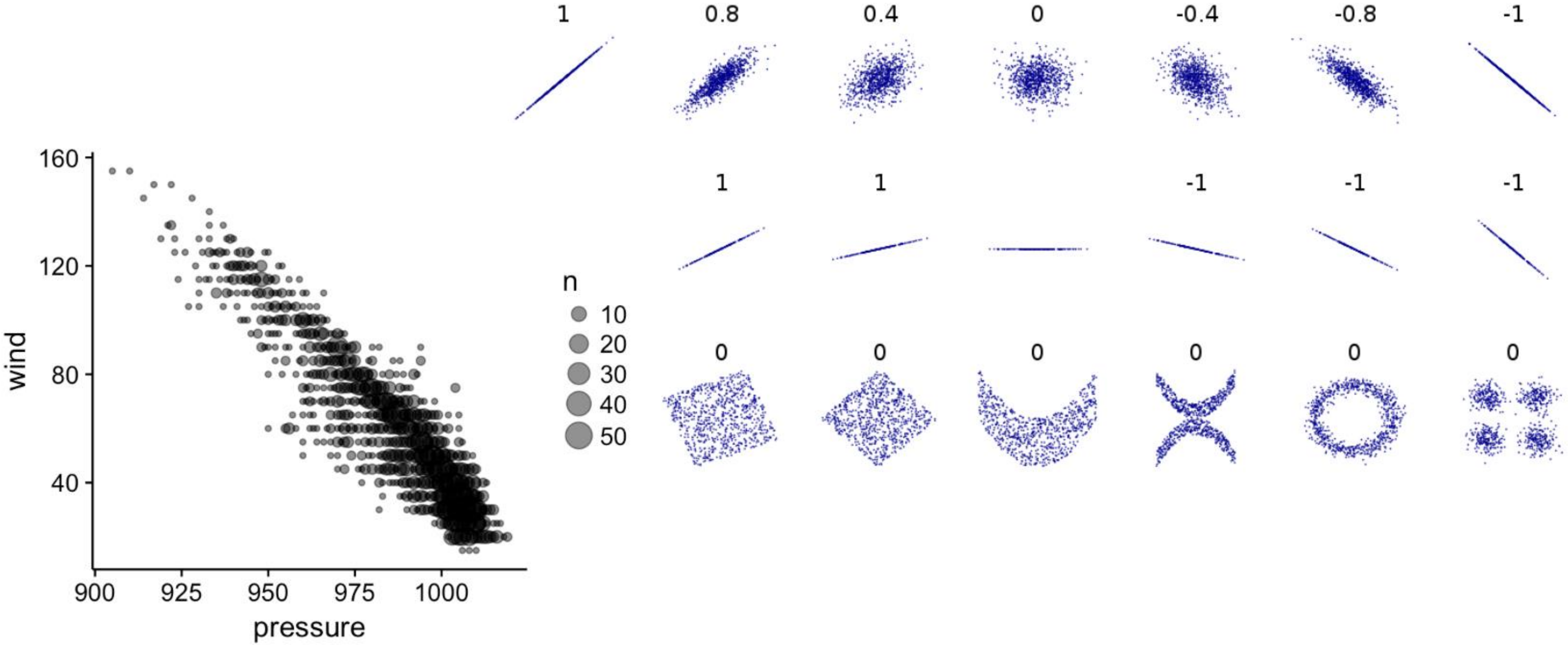


Step 3

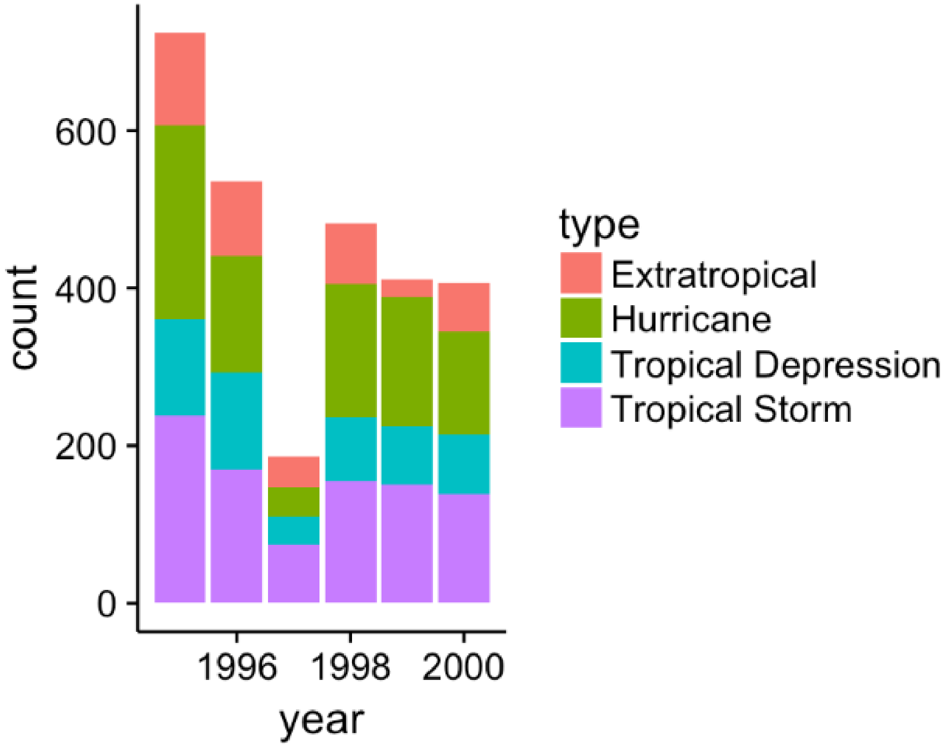
Find association between variables

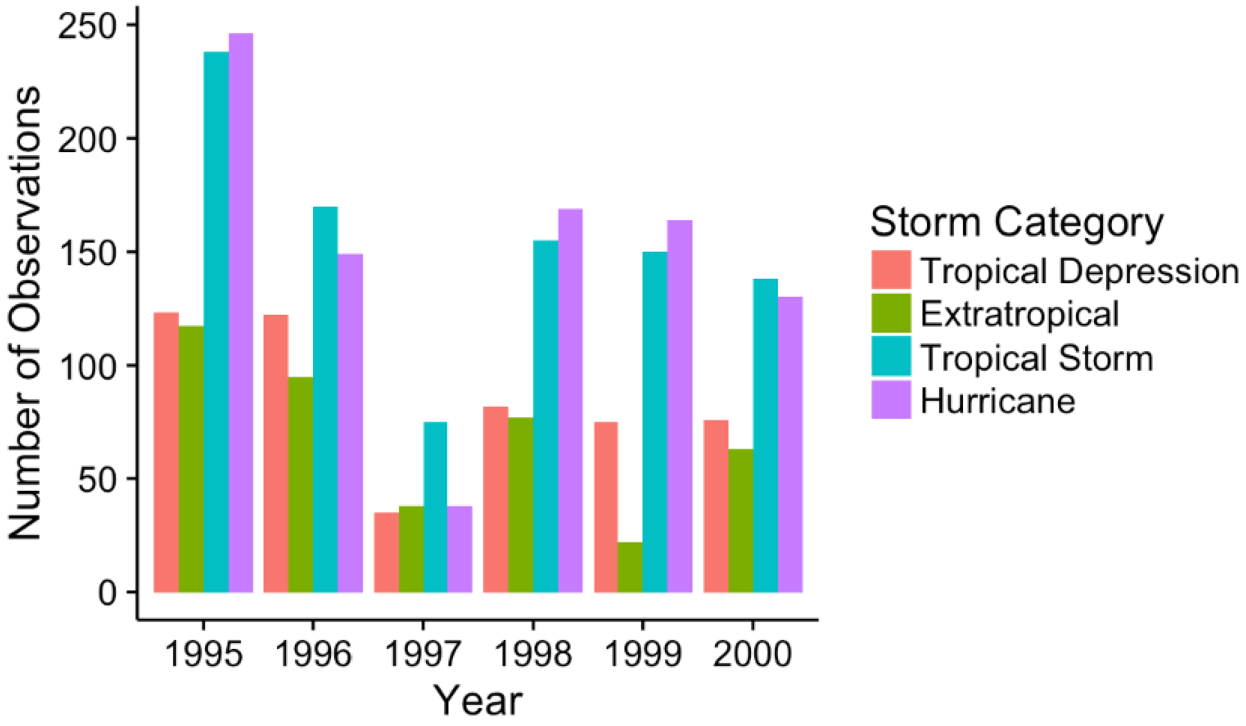
IBM **Developer**

Associations between numerical variables

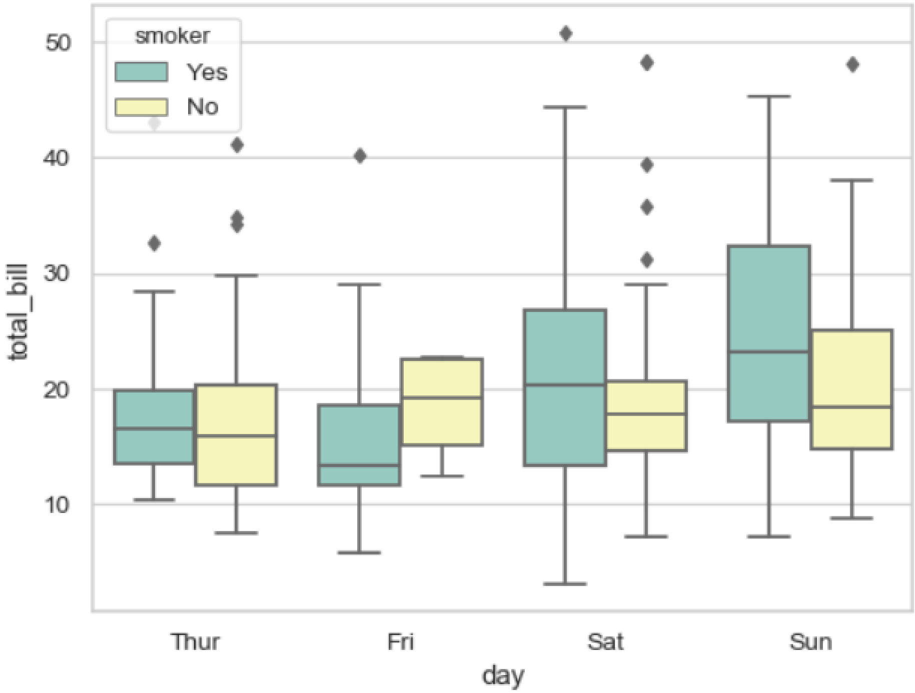


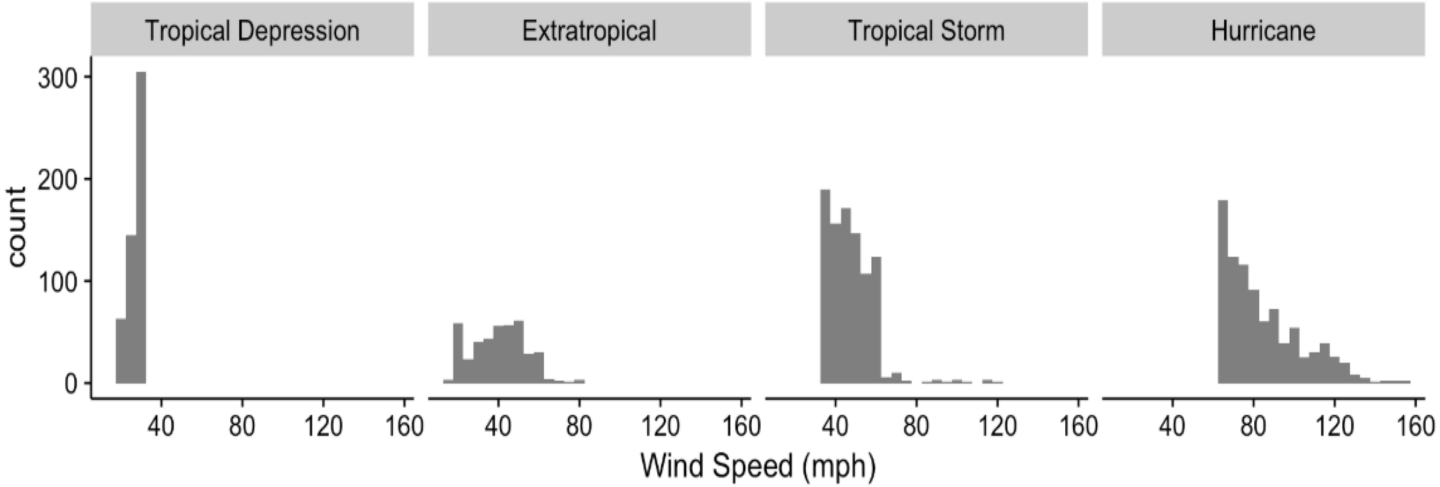
Associations between categorical variables





Associations between numerical-categorical variables

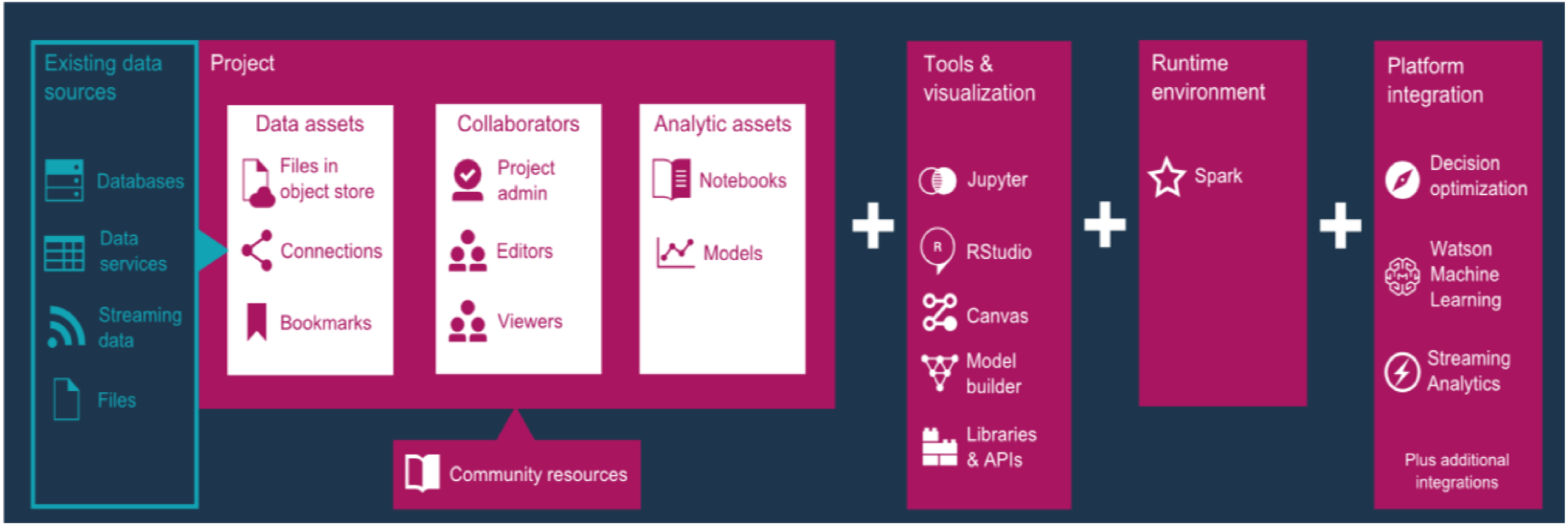




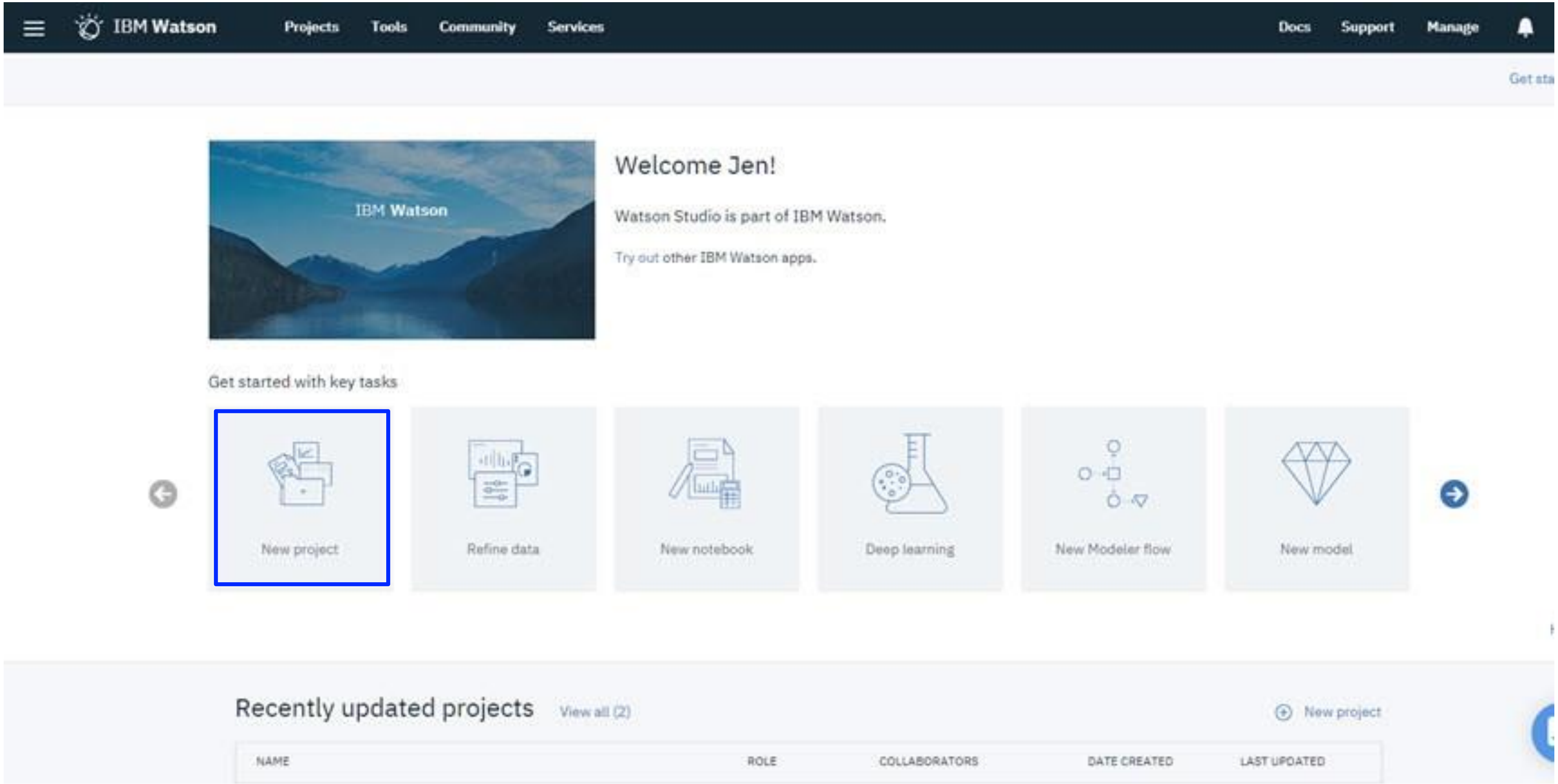
Watson Studio

Exploration

Watson Studio



Watson Studio Dashboard



Welcome Jen!

Watson Studio is part of IBM Watson.

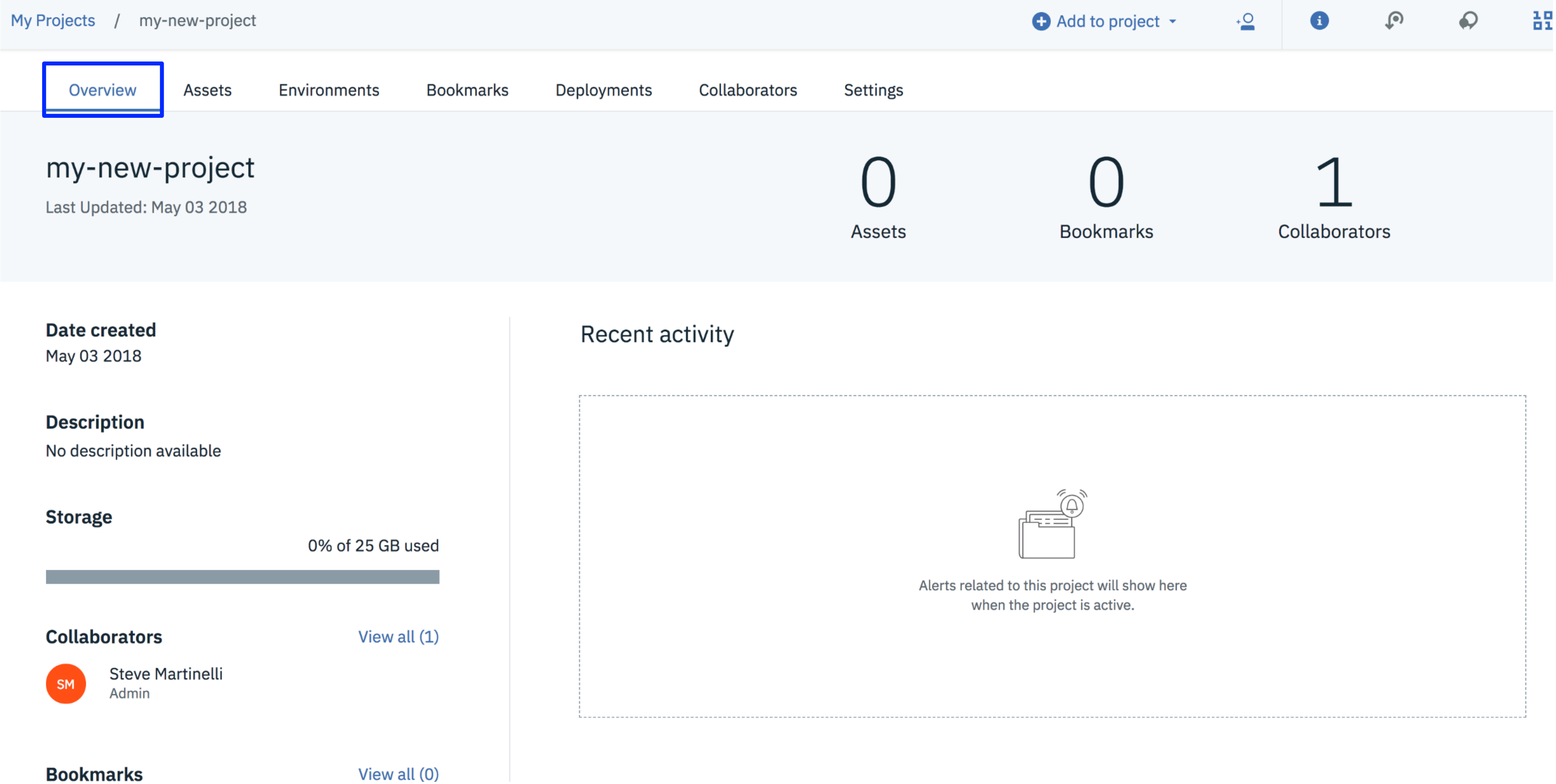
Try out other IBM Watson apps.

Get started with key tasks

Recently updated projects

NAME	ROLE	COLLABORATORS	DATE CREATED	LAST UPDATED

Overview Page



my-new-project

Last Updated: May 03 2018

0 Assets

0 Bookmarks

1 Collaborators

Date created

May 03 2018

Description

No description available

Storage

0% of 25 GB used

Collaborators View all (1)

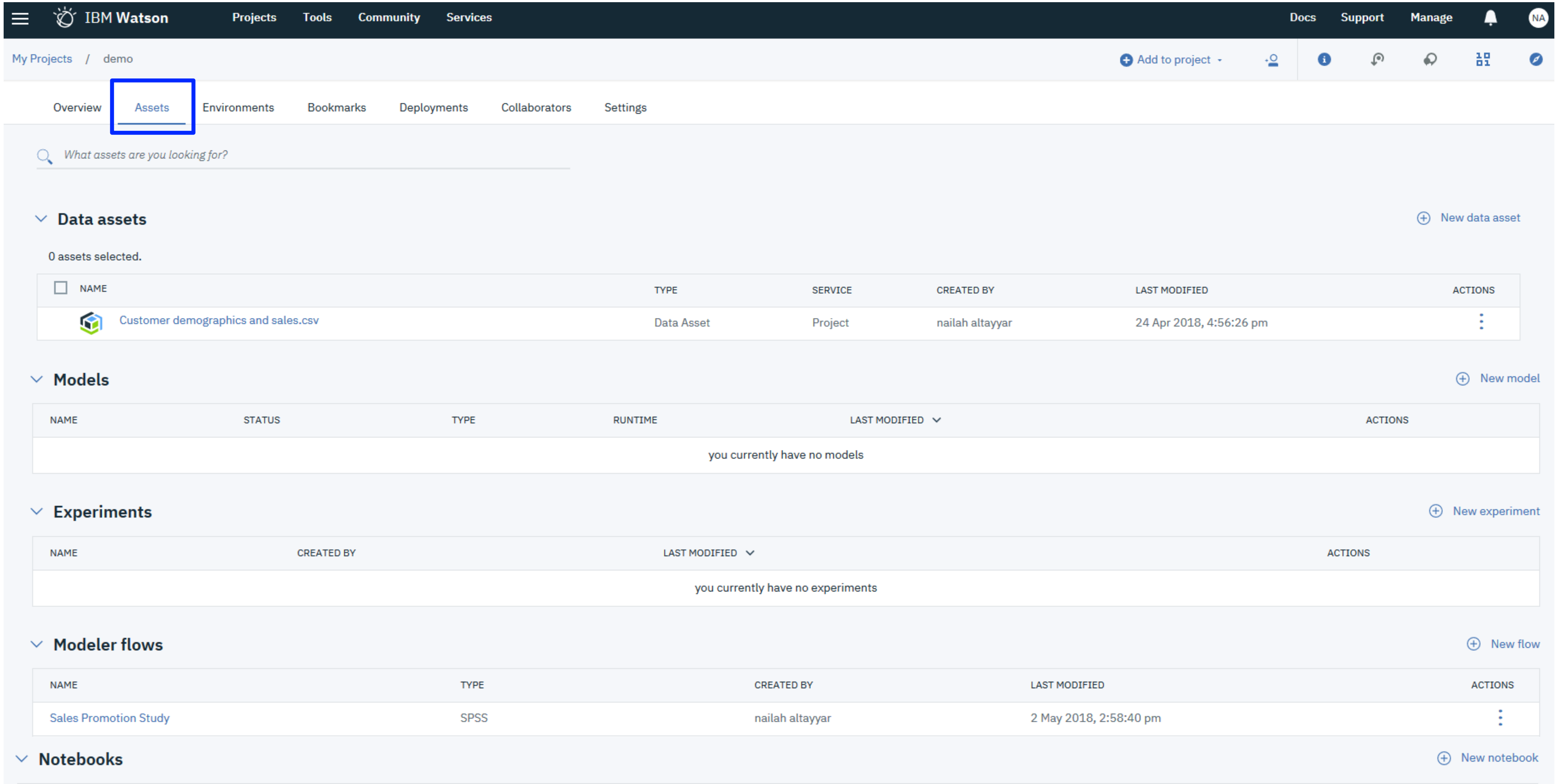
SM Steve Martinelli
Admin

Bookmarks View all (0)

Recent activity

Alerts related to this project will show here when the project is active.

Assets Page



Data assets

New data asset

0 assets selected.

NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
Customer demographics and sales.csv	Data Asset	Project	nailah altayyar	24 Apr 2018, 4:56:26 pm	⋮

Models

New model

NAME	STATUS	TYPE	RUNTIME	LAST MODIFIED	ACTIONS
you currently have no models					

Experiments

New experiment

NAME	CREATED BY	LAST MODIFIED	ACTIONS
you currently have no experiments			

Modeler flows

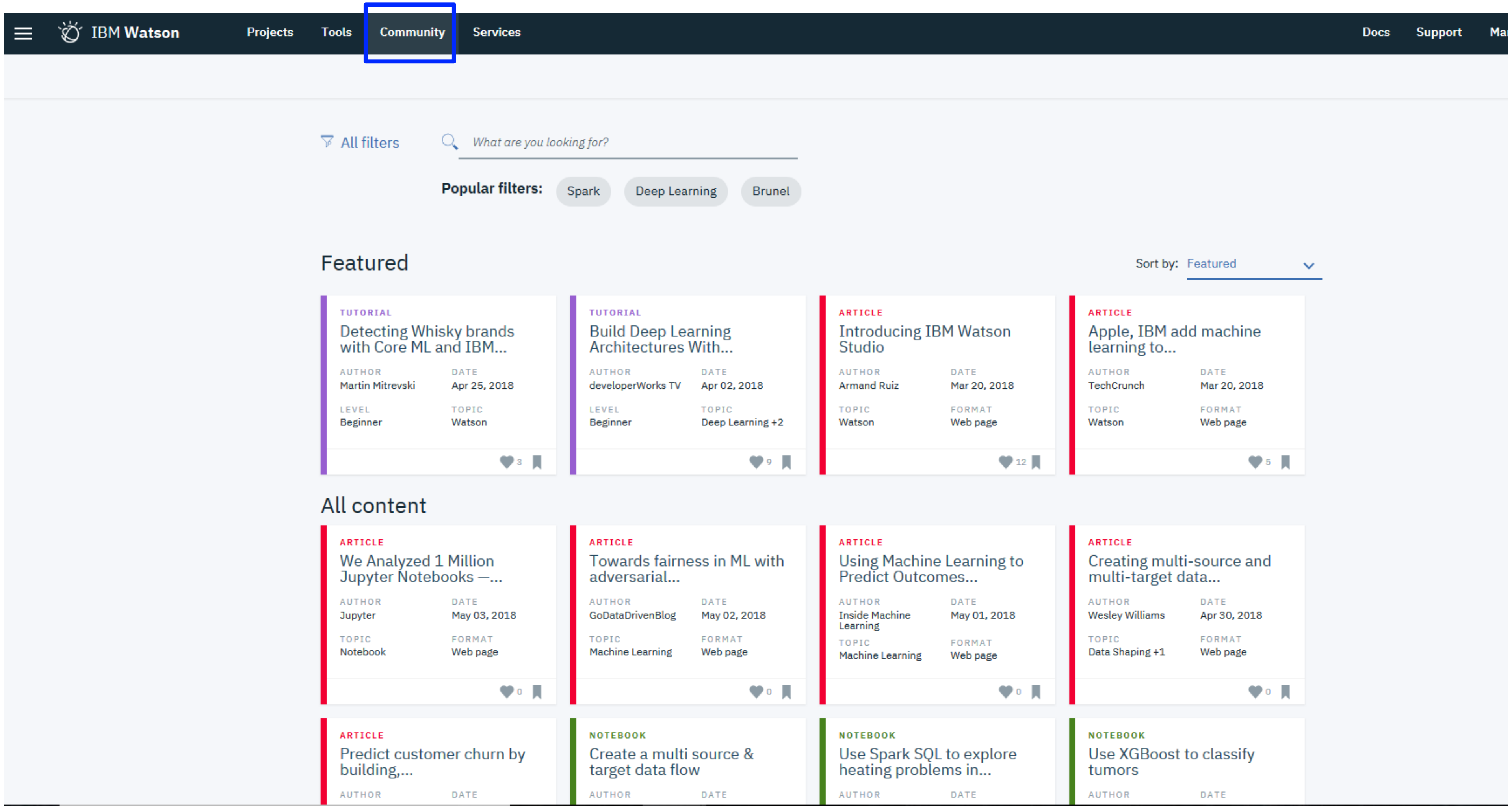
New flow

NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
Sales Promotion Study	SPSS	nailah altayyar	2 May 2018, 2:58:40 pm	⋮

Notebooks

New notebook

Community



IBM Watson Projects Tools **Community** Services Docs Support Ma

All filters What are you looking for?

Popular filters: Spark Deep Learning Brunel

Featured

Sort by: Featured

TUTORIAL
Detecting Whisky brands with Core ML and IBM...
AUTHOR Martin Mitrevski DATE Apr 25, 2018
LEVEL Beginner TOPIC Watson
♥ 3

TUTORIAL
Build Deep Learning Architectures With...
AUTHOR developerWorks TV DATE Apr 02, 2018
LEVEL Beginner TOPIC Deep Learning +2
♥ 9

ARTICLE
Introducing IBM Watson Studio
AUTHOR Armand Ruiz DATE Mar 20, 2018
TOPIC Watson FORMAT Web page
♥ 12

ARTICLE
Apple, IBM add machine learning to...
AUTHOR TechCrunch DATE Mar 20, 2018
TOPIC Watson FORMAT Web page
♥ 5

All content

ARTICLE
We Analyzed 1 Million Jupyter Notebooks —...
AUTHOR Jupyter DATE May 03, 2018
TOPIC Notebook FORMAT Web page
♥ 0

ARTICLE
Towards fairness in ML with adversarial...
AUTHOR GoDataDrivenBlog DATE May 02, 2018
TOPIC Machine Learning FORMAT Web page
♥ 0

ARTICLE
Using Machine Learning to Predict Outcomes...
AUTHOR Inside Machine Learning DATE May 01, 2018
TOPIC Machine Learning FORMAT Web page
♥ 0

ARTICLE
Creating multi-source and multi-target data...
AUTHOR Wesley Williams DATE Apr 30, 2018
TOPIC Data Shaping +1 FORMAT Web page
♥ 0

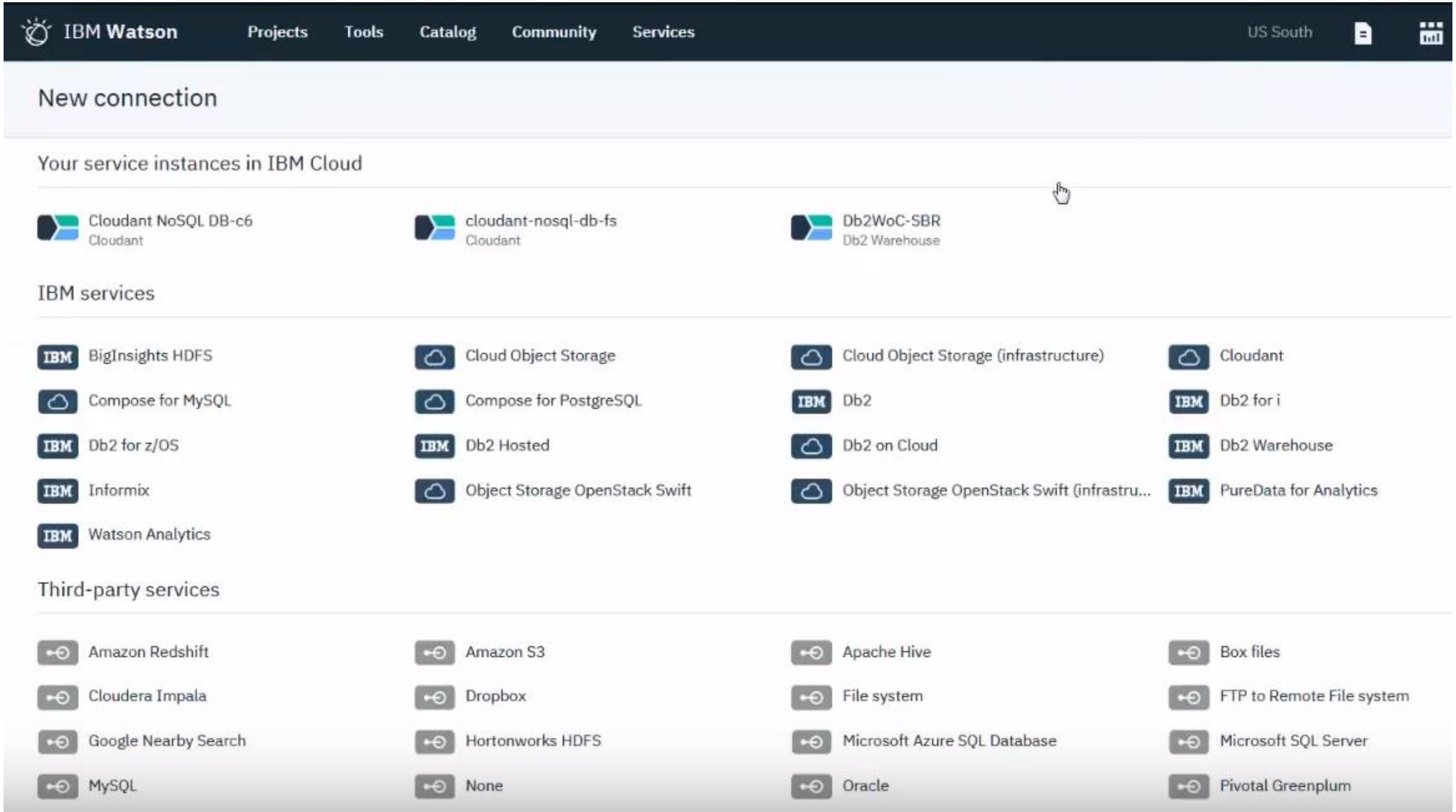
ARTICLE
Predict customer churn by building,...
AUTHOR DATE

NOTEBOOK
Create a multi source & target data flow
AUTHOR DATE

NOTEBOOK
Use Spark SQL to explore heating problems in...
AUTHOR DATE

NOTEBOOK
Use XGBoost to classify tumors
AUTHOR DATE

Connections



New connection

Your service instances in IBM Cloud

- Cloudant NoSQL DB-c6 — Cloudant
- cloudant-nosql-db-fs — Cloudant
- Db2WoC-SBR — Db2 Warehouse

IBM services

- BigInsights HDFS
- Cloud Object Storage
- Cloud Object Storage (infrastructure)
- Cloudant
- Compose for MySQL
- Compose for PostgreSQL
- Db2
- Db2 for i
- Db2 for z/OS
- Db2 Hosted
- Db2 on Cloud
- Db2 Warehouse
- Informix
- Object Storage OpenStack Swift
- Object Storage OpenStack Swift (infrastru...
- PureData for Analytics
- Watson Analytics

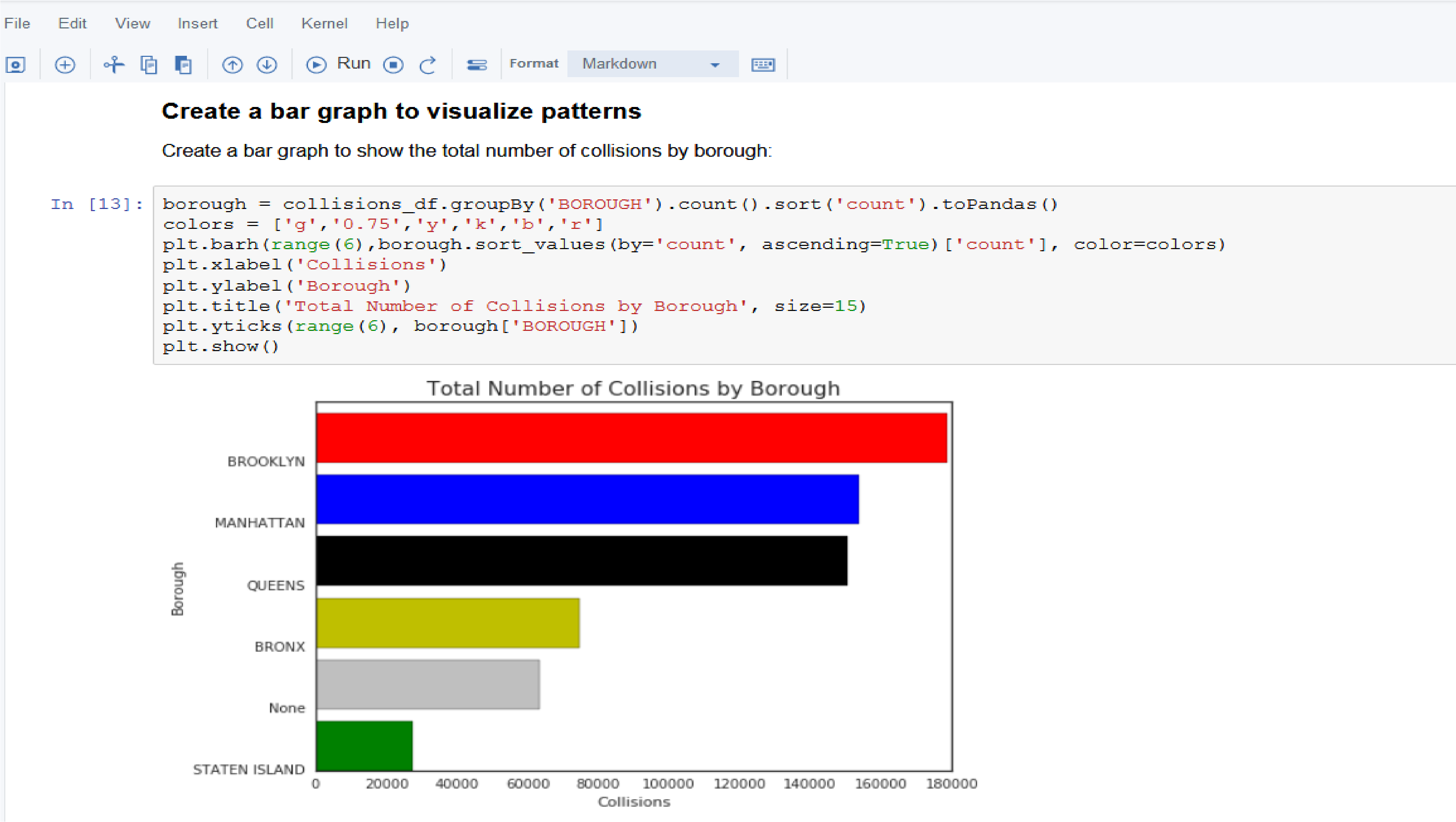
Third-party services

- Amazon Redshift
- Amazon S3
- Apache Hive
- Box files
- Cloudera Impala
- Dropbox
- File system
- FTP to Remote File system
- Google Nearby Search
- Hortonworks HDFS
- Microsoft Azure SQL Database
- Microsoft SQL Server
- MySQL
- None
- Oracle
- Pivotal Greenplum

In Watson Studio, after you set up a project and add data to it, you can start analyzing and visualizing your data:

With Code
Or Without Code

Jupyter Notebook



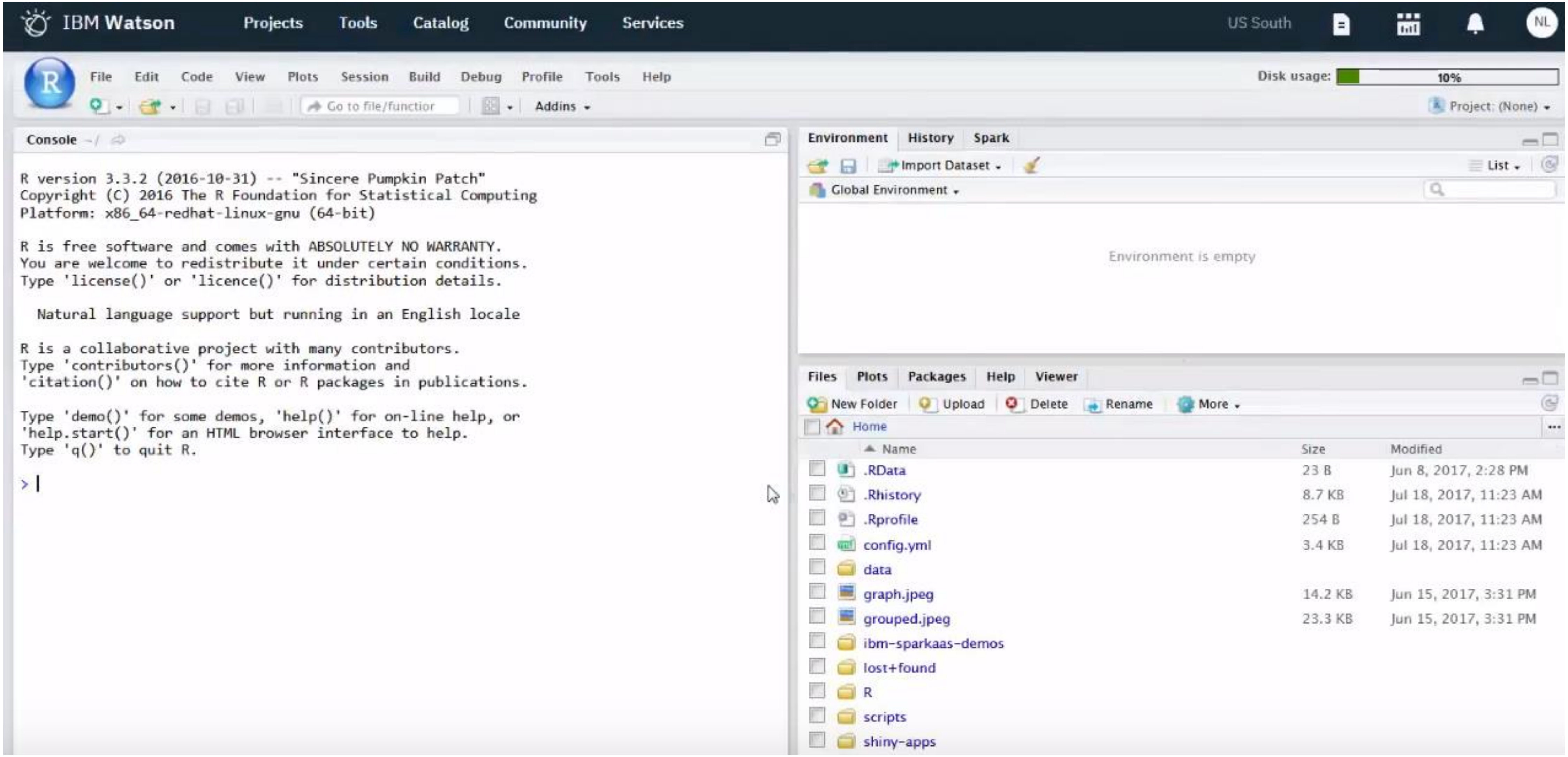
Create a bar graph to visualize patterns

Create a bar graph to show the total number of collisions by borough:

In [13]:

```
borough = collisions_df.groupBy('BOROUGH').count().sort('count').toPandas()
colors = ['g','0.75','y','k','b','r']
plt.barh(range(6),borough.sort_values(by='count', ascending=True)['count'], color=colors)
plt.xlabel('Collisions')
plt.ylabel('Borough')
plt.title('Total Number of Collisions by Borough', size=15)
plt.yticks(range(6), borough['BOROUGH'])
plt.show()
```

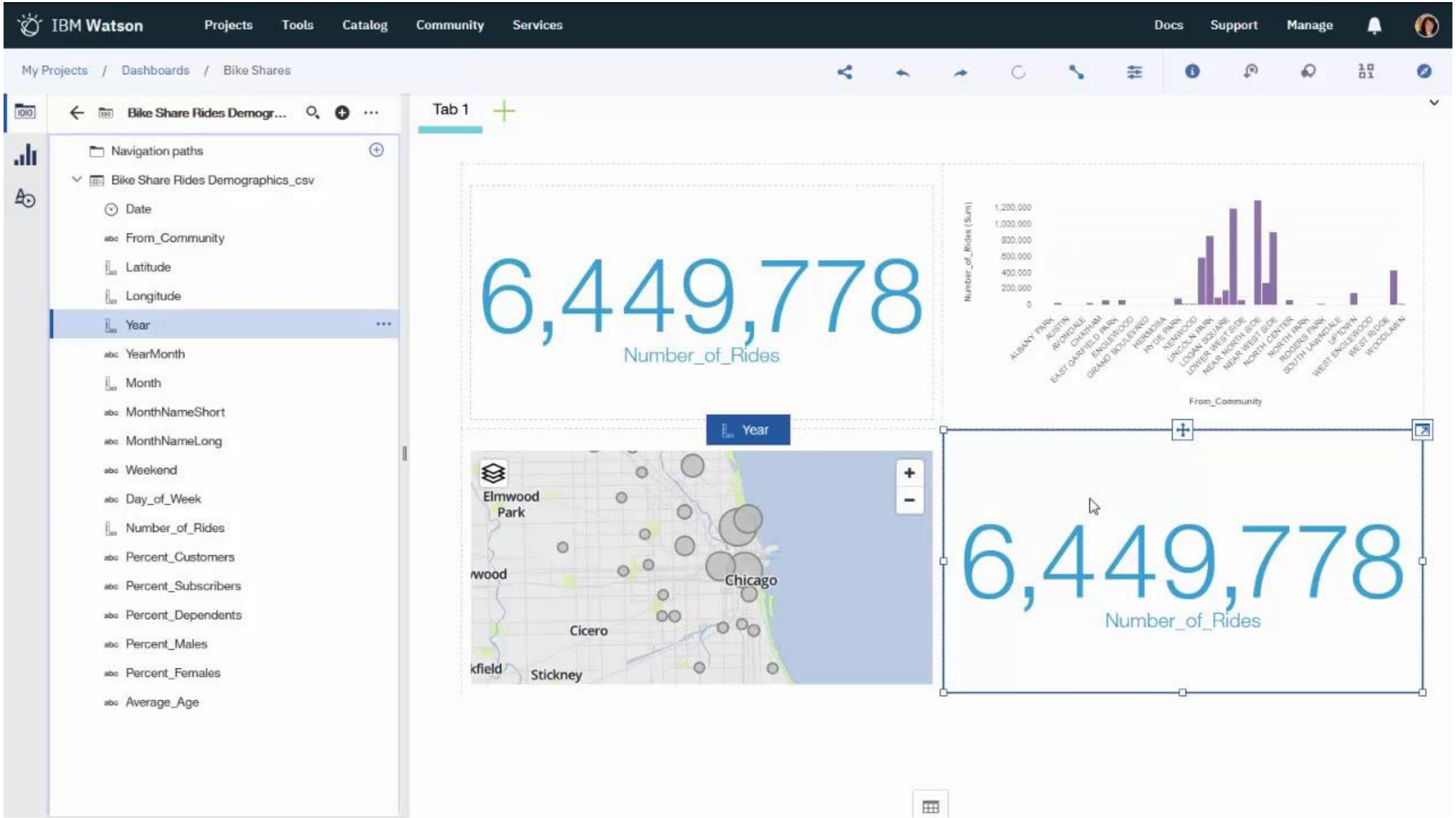
RStudio



```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Name	Size	Modified
.RData	23 B	Jun 8, 2017, 2:28 PM
.Rhistory	8.7 KB	Jul 18, 2017, 11:23 AM
.Rprofile	254 B	Jul 18, 2017, 11:23 AM
config.yml	3.4 KB	Jul 18, 2017, 11:23 AM
data		
graph.jpeg	14.2 KB	Jun 15, 2017, 3:31 PM
grouped.jpeg	23.3 KB	Jun 15, 2017, 3:31 PM
ibm-sparkaas-demos		
lost+found		
R		
scripts		
shiny-apps		

Dashboard


IBM Watson
Projects
Tools
Catalog
Community
Services
Docs
Support
Manage
My Projects / Dashboards / Bike Shares
Bike Share Rides Demogr...
Tab 1
Navigation paths
Bike Share Rides Demographics_csv
Date
From_Community
Latitude
Longitude
Year
YearMonth
Month
MonthNameShort
MonthNameLong
Weekend
Day_of_Week
Number_of_Rides
Percent_Customers
Percent_Subscribers
Percent_Dependents
Percent_Males
Percent_Females
Average_Age
6,449,778
Number_of_Rides
Number_of_Rides (Sum)
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
ALBANY PARK
AUSTIN
AVONDALE
CHATHAM
EAST GARFIELD PARK
ENGLEWOOD
GRAND BOULEVARD
HERMOSA
HYDE PARK
KENWOOD
LINCOLN PARK
LOGAN SQUARE
LOWER WEST SIDE
NEAR NORTH SIDE
NEAR WEST SIDE
NORTH CENTER
NORTH PARK
ROGERS PARK
SOUTH LAWNDALE
UPTOWN
WEST ENGLEWOOD
WEST RIDGE
WOODLAWN
From_Community
Year
Elmwood Park
wood
Chicago
Cicero
kfield
Stickney
+
−
6,449,778
Number_of_Rides

Titanic Dataset

The sinking of the Titanic is one of the most famous shipwrecks in history. The Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew.

This sensational tragedy shocked the international community and led to better safety regulations for ships.

Data Description

Description:

Information about the passenger of titanic, 891 case.

Format

survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name of passenger
Gender	Gender
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Source https://www.kaggle.com/c/titanic/data

Please, Sign Up for IBM Cloud (US)

https://ibm.biz/BdYpAP

GitHub

https://github.com/DevExCodeHub/DataScienceSeries-Ep1

Wi-Fi Password: makeithappen