Introduction to Data Analytics

PROFESSIONAL CERTIFICATE IBM Data Analyst

Agenda

- Analyzing and Mining Data
- Communicating Data Analysis Findings

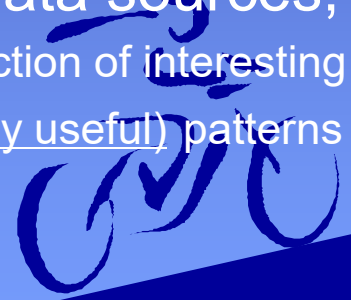
Analyzing And Mining Data

Statistical Analysis

- **Statistical** knowledge helps you use the proper methods to collect the data, employ the correct analyses, and effectively present the results. **Statistics** is a crucial process behind how we make discoveries in science, make decisions based on data, and make predictions.
- **Statistical Analysis can be:**
- Descriptive; that which provides a summary of what the data represents. Common measures include Central Tendency, Dispersion, and Skewness.

Data Mining

- Data mining is also called *knowledge discovery and data mining* (KDD)
- Data mining is
  - extraction of useful patterns from data sources, e.g., databases, texts, web, image. Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Patterns must be:
  - valid, novel, potentially useful, understandable



Example of discovered patterns

- Association rules:

  "80% of customers who buy *cheese* and *milk* also buy *bread*, and 5% of customers buy all of them together"

Potential Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Data Mining Applications

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction.

Telecommunication Industry

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics.

Ex: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, surveys …
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.,
    - E.g. Most customers with income level 60k – 80k with food expenses $600 - $800 a month live in that area
  - Determine customer purchasing patterns over time
    - E.g. Customers who are between 20 and 29 years old, with income of 20k – 29k usually buy this type of CD player
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
  - E.g. Customers who buy computer A usually buy software B

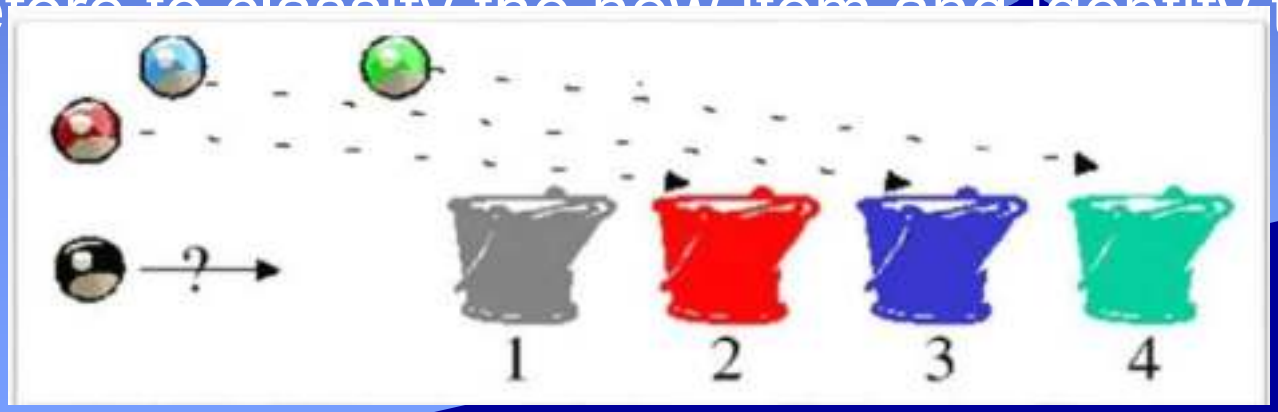
- Customer requirement analysis
  - Identify the best products for different customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
    - E.g. Summarize all transactions of the first quarter from three different branches
      Summarize all transactions of last year from a particular branch
      Summarize all transactions of a particular product
  - Statistical summary information
    - E.g. What is the average age for customers who buy product A?
- Fraud detection
  - Find outliers of unusual transactions
- Financial planning
  - Summarize and compare the resources and spending

Data Mining Techniques

- Classification
- Clustering
- Regression
- Association Rules/ Affinity Grouping
- Anomaly or outlier detection
- Sequential Patterns
- Decision Trees

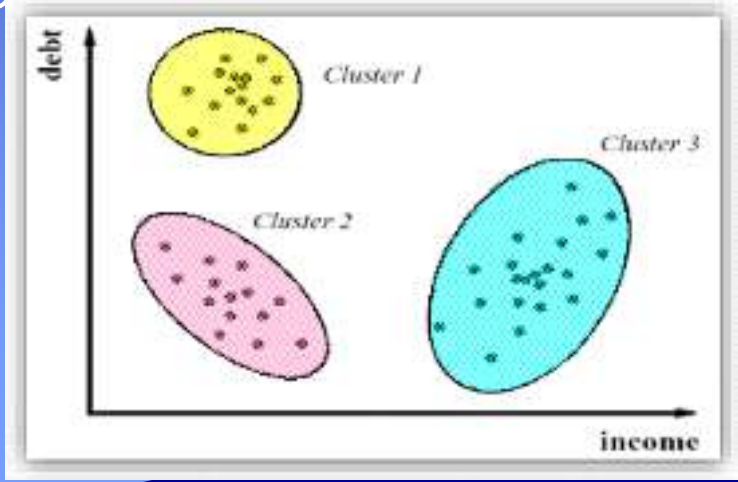
Classification

- Classification is the process of predicting the class of a new item.
- Therefore to classify the new item and identify to which



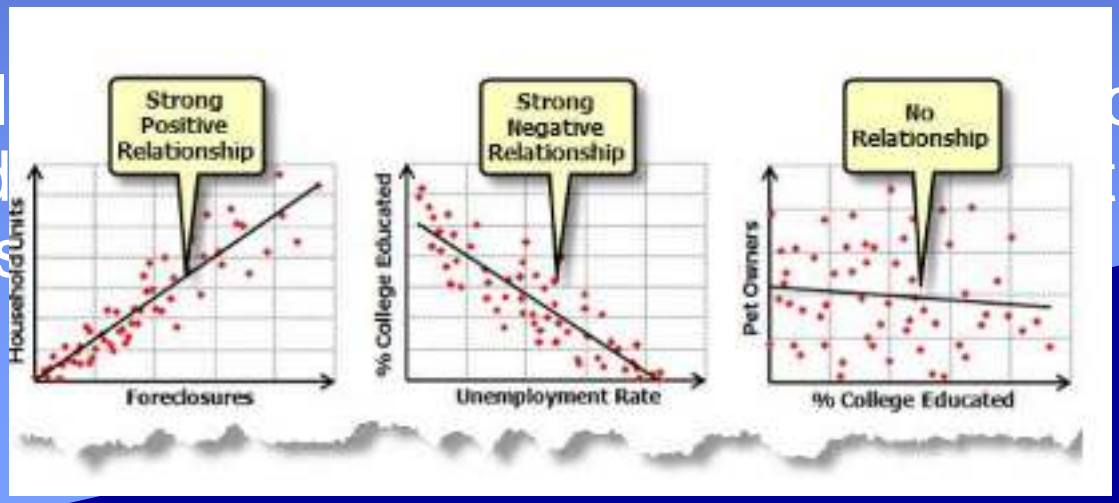
Clustering

- Group Data into Clusters
- Similar data is grouped in the same cluster
- Dissimilar data is grouped in the same cluster



Regression Analysis

- “Regression deals with the prediction of a value, rather than a class.”
- Regression is a data mining function that predicts a number
- For exampl d to predict child , and other factors



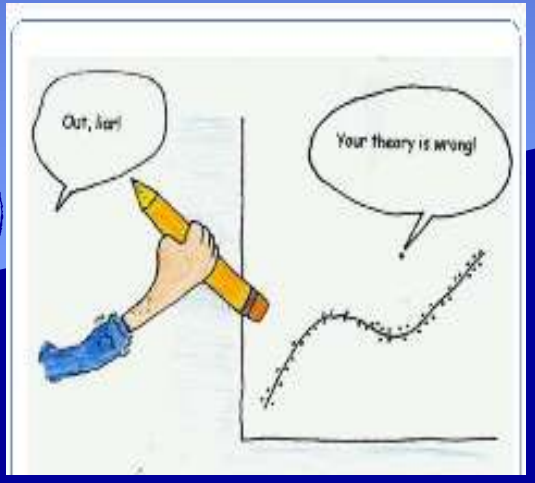
Association Rules / Affinity grouping

- “An association algorithm creates rules that describe how often events have occurred together.”
- Example: When a customer buys a Computer, then 90% of the time they will buy softwares

Outlier Detection

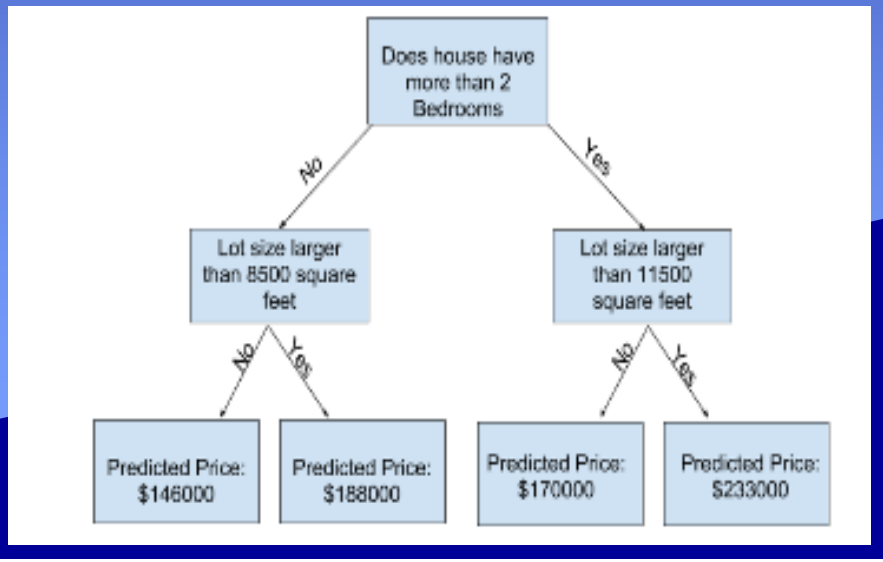
- This technique is used for identifying unusual or suspicious cases that deviate from the projected pattern or expected norm. The applications of

IP address	User ID	Account Number	Age	Shipping Address	Transaction Date	Transaction Time	Transaction Value	Product Category	Units Purchased
3.56.123.0	johns	25671147	32	1542, Orchid Lane, WA 98706, US	15-5-20	15:00:05	$131.58	Clothing	1
3.56.123.0	johns	25671147	32	1542, Orchid Lane, WA 98706, US	10-6-20	10:23:50	$79.33	Electronics	2
3.56.123.0	johns	25671147	32	1542, Orchid Lane, WA 98706, US	1-6-20	07:12:45		Home Décor	1
1.180.52.7	johns	25671147	32	In-store	5-6-20	01:11:10	$2,009.99	Electronics	10
	johns	25671147	32	In-store	2020-06-03	01:15:12	$4,131.00	Electronics	15
1.180.52.7	johns	25671147	32	P.O. Box 1049	09-06-2020	01:22:24	$8,010.50	Tools	20
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	15 May 2020	17:03:08	$254.20	Furniture	1
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	18 May 2020	19:12:45	$141.00	Kitchen Supplies	3
	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	01 June 2020	17:34:15	$157.25	Car Spares	2
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	15 June 2020	18:01:10	$80.00	Kitchen Supplies	1
172.165.10.1	ellerd	11568528		P.O. Box 1322	07 June 2020	15:55:12	$99.99	Clothing	1
172.165.10.1	ellerd	11568528		P.O. Box 1322	08 June 2020	17:15:30	$53.19	Beauty	1
1.167.255.10	ellerd	11568528		P.O. Box 5401	02 July 2020	00:05:10	$4,895.00	Laptop	1



Decision Trees

Decision tree learning or **induction of decision trees** is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).



Sequential pattern

- **Sequential pattern mining** is a topic of data miningconcerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence.It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity.
- Example, "if a {customer buys a car}, he or she is likely to {buy insurance} within 1 week",

Tools for Data Mining

- Spreadsheets
- Python
- IBM SPSS
- R- Language
- IBM Watson Studio
- SAS

Communicating Data Analysis Findings

Data Analysis Process

Who is my audience?

What is important to them?

What will help them trust me?

Structure Your presentation

- Reference your data
- State your assumption
- Organize your presentation
- Identify the best formats for presenting your data

Viewpoints: Storytelling in Data Analysis

- Storytelling in Data Analysis

Introduction to Visualization and Dashboarding Software

- List some of the most commonly used data visualization software?

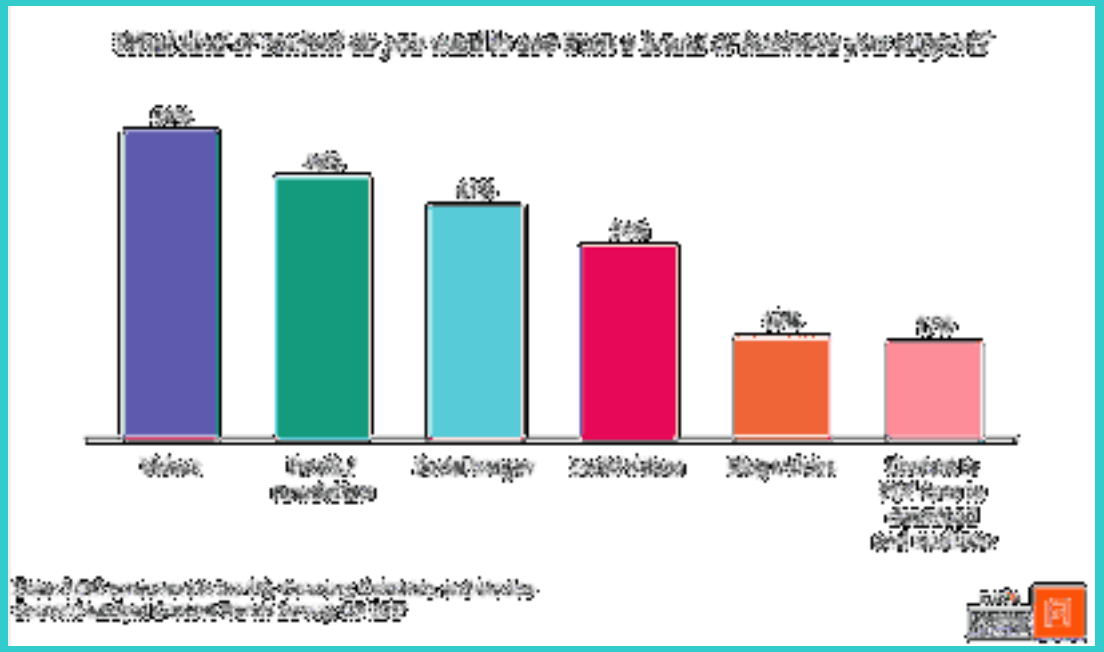
Introduction to Visualization and Dashboarding Software

- List some of the most commonly used data visualization software?

- Spreadsheets
- Jupyter Notebook
- Python libraries, like what?!
- R-Studio and R-Shiny,
- IBM Cognos Analytics
- Tableau and Microsoft Power BI.

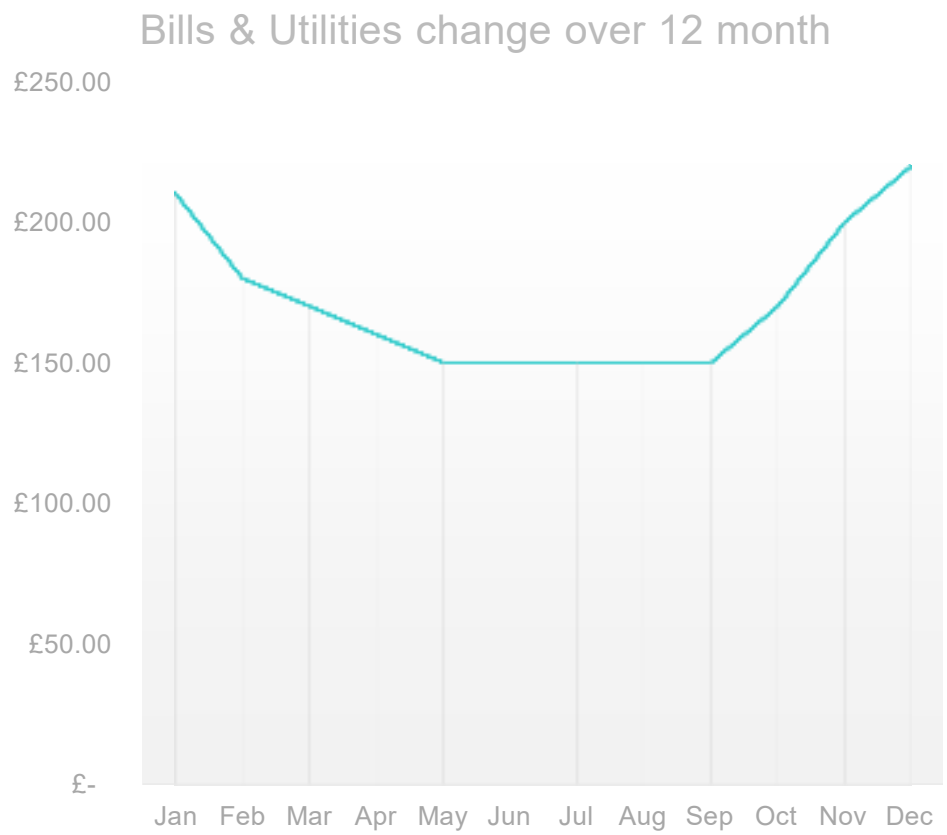
Activity

- What is the story?



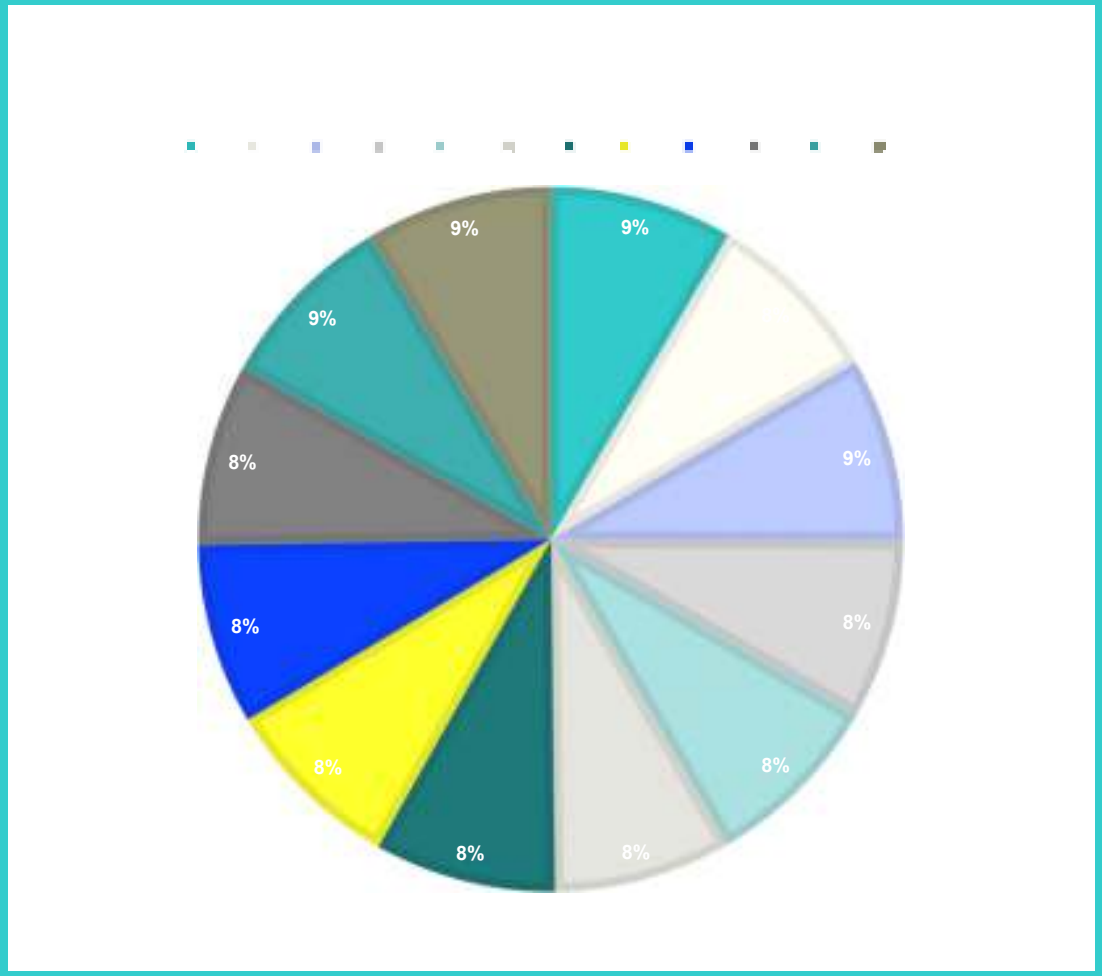
Activity

- What is the story?



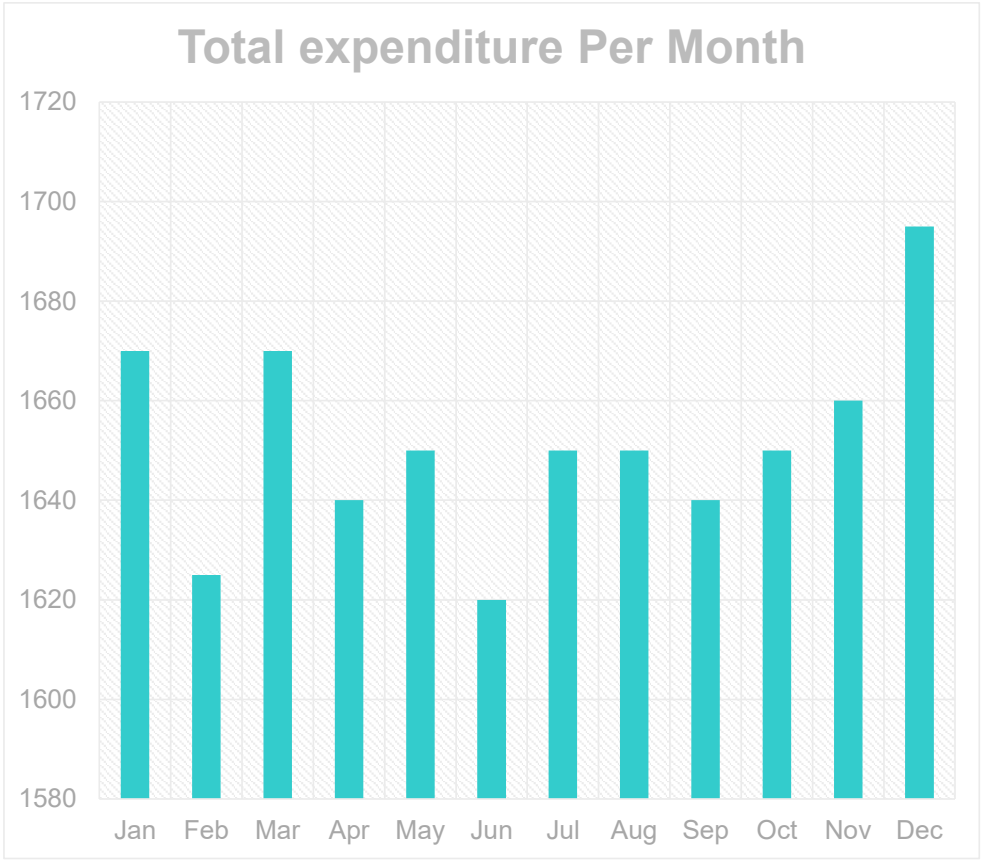
Activity

- What is the story?



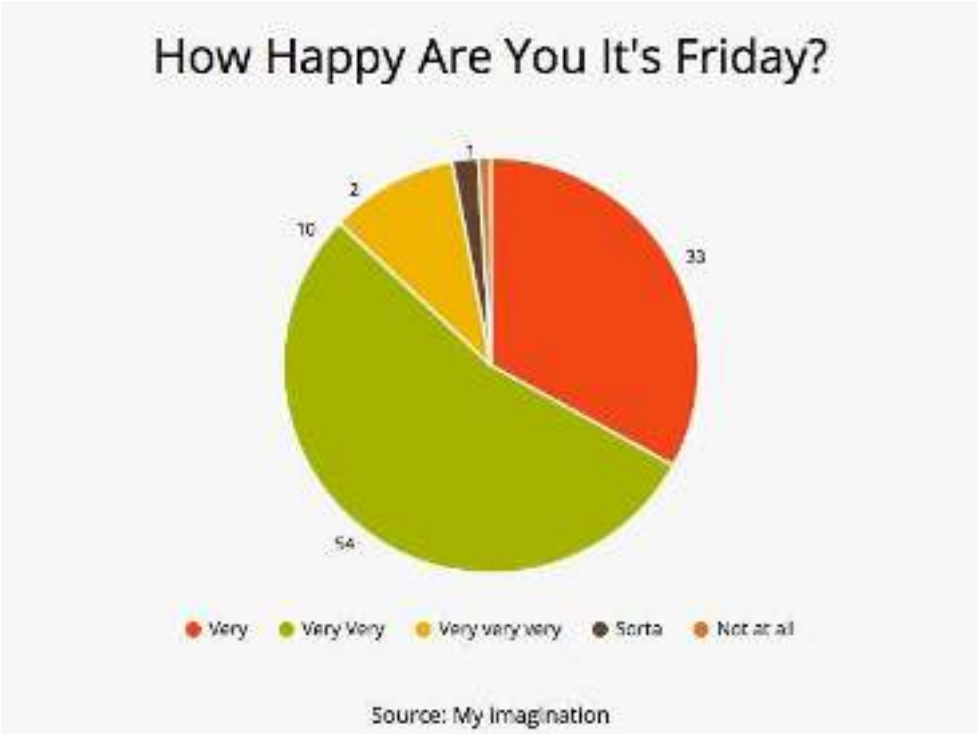
Activity

- What is the story?



Activity

- What is the story?



Activity

- What is the story?