Universidad Politécnica de Yucatán

Major: Data Engineering

Group: 9-A

Subject: Business Intelligence

Homework: 3.1 Data Visualization

Teacher: Didier Omar Gamboa Angulo

Students:

Sergio Johanan Barrera Chan

Ariel Joel Buenfil Góngora

Damaris Yuselin Dzul

Diego Monroy Minero

Alan Alberto Valbuena

Due Date: Thursday, October 28th, 2025

1. Opening Vignette: SiriusXM

Summary

SiriusXM, a major satellite radio company, faced challenges due to shifting customer demographics (skewing younger with less discretionary income), the need to convert second owners of SiriusXM-enabled cars to paying customers, and leveraging a new hybrid satellite/wireless delivery capability acquired from Agero. Their proposed solution was to transform into a high-performance, **data-driven marketing organization** focusing on personalized customer interactions, enabling integrated data and advanced analytics, and achieving an integrated and consistent internal point of view.

The company brought its marketing capabilities in-house and partnered with Teradata. Key implementation steps included improving data cleanliness, expanding the data warehouse for integrated marketing analytics, developing new segmentation and scoring models, and adopting a modern marketing platform. As a result, campaign cycle times were massively reduced (from 4 days to near real-time), enabling quicker analysis, real-time modeling, and more targeted offers at the household and device levels.

Discussion Questions

1. What does SiriusXM do? In what type of market does it conduct its business?

- SiriusXM Radio is a **satellite radio powerhouse**, providing a wide range of music, sports, news, talk, and entertainment stations. It is the largest radio company in the world.
- It operates in the highly competitive **entertainment and media market**, specifically satellite radio, facing challenges from changing customer demographics, technology, and competition.

2. What were the challenges? Comment on both technology and data-related challenges.

- **Market/Demographic Challenges:** Changing customer demographics (skewing younger with less discretionary income), and the need to identify and convert second-hand car owners to subscribers.
- **Technology/Data-related Challenges:** The existing third-party marketing platform couldn't support SiriusXM's ambitions. The company needed to improve **data cleanliness** through better master data management and governance, gain **integrated data** capabilities for advanced analytics and a coherent **360-degree view** of all consumers, and resolve **data latency issues** to deliver real-time offer management.

3. What were the proposed solutions?

- The proposed solution was to become a **high-performance, data-driven marketing organization**. This was based on three tenets:
    - **Personalized interactions** instead of mass marketing.

    - Leveraging **integrated data, advanced analytics**, and integrated marketing platforms.
    - Ensuring an **integrated and consistent point of view** across the business and technology sides.

4. How did they implement the proposed solutions? Did they face any implementation challenges?

- They implemented the solution by first deciding to **bring marketing capabilities in-house** and partnering with **Teradata**. They followed a multi-step path:
    - Improving **data cleanliness** (master data management and governance).
    - Bringing **marketing analytics in-house** and expanding the data warehouse (Integrated Data Warehouse - IDW).
    - Developing new **segmentation and scoring models** to run *in-database*.
    - Adopting the Teradata Integrated Marketing Cloud (Customer Interaction Manager, Digital Messaging Center) for campaign development and **real-time offer management** across channels.
- The document does not explicitly list "implementation challenges" they faced, but it highlights that getting **data hygiene** (cleanliness) correct was a **necessary first step** and that the company is in the midst of a transformational, 5-year process, suggesting a large, complex undertaking.

5. What were the results and benefits? Were they worth the effort/investment?

- The results and benefits included:
    - **Massive reductions in campaign cycle times**: from 4 days to near real-time.
    - Ability to create **more targeted offers** at the household, consumer, and device levels.
    - **Real-time modeling and scoring** to increase marketing intelligence.
    - **Closed-loop visibility** for analysts to support multistage dialogues and in-campaign modifications.
    - **More tailored and effective marketing** is typically more cost-efficient.
- The effort and investment appear to be **worth it**, as the company is gaining significant competitive advantages, driving efficiency, and seeing tangible improvements in marketing speed and effectiveness.

6. Can you think of other companies facing similar challenges that can potentially benefit from similar data-driven marketing solutions?

- **Telecommunications Companies (e.g., mobile carriers):** They face high churn rates (attrition) and need to cross-sell/up-sell services (data, phone, internet) across devices and customer life cycles.

- **Subscription Box Services/E-commerce:** They need to personalize product recommendations, manage inventory based on predicted demand, and re-engage customers whose subscriptions are about to expire.
- **Insurance Companies:** They need to target personalized policy offers to customers based on demographics/history, and cross-sell different types of insurance (auto, home, life) to the same household.

2. Readiness Level of Data for Analytics Study

Data must be **analytics ready**—relevant to the problem, meeting quality/quantity requirements, and having a proper structure.

Metric	Definition	Example	Evaluation Method
Data source reliability	The originality and appropriateness of the storage medium where the data is obtained (Do we have confidence in this source?).	Seeking the original source of customer survey data rather than a fifth-hand summary table.	Comparing data moved through multiple stops to the original source to check for unintentional drops or reformatting.
Data content accuracy	Data are correct and a good match for the analytics problem (Do we have the right data for the job?).	A customer's contact information in the database matches what the customer originally provided.	Checking if recorded values represent what was intended or defined by the original source.

Data accessibility	The data are easily and readily obtainable (Can we easily get to the data when needed?).	Being able to merge and transform customer and sales data that is stored across multiple databases or mediums like data lakes.	Assessing the complexity of accessing data stored in disparate locations or modern infrastructures like Hadoop.
Data security & privacy	Data is secured to only allow authorized people to access it and prevent unauthorized access.	Implementing systems that comply with laws like HIPAA to safeguard patient health records.	Verifying systems are in place to only allow those with the authority and need to access the data.
Data richness	All the required data elements (variables) are included in the data set to portray a rich enough dimensionality of the subject matter.	A customer data set includes not only purchase history but also demographic and social media activity to build a highly predictive model.	Checking if the information content is complete (or near complete) to build a predictive or prescriptive model.

Data consistency	Data are accurately collected and combined/merged, ensuring dimensional information pertains to the same subject.	Ensuring that after merging, one patient record does not accidentally contain variables mixed up from a different patient's record.	Checking variables from disparate sources to ensure they pertain to the same subject after integration/merging.
Data currency/timeliness	Data should be up-to-date and recorded at or near the time of the event or observation.	Sales data used for next month's forecast must be the most recent daily or weekly totals, not data from 3 months ago.	Assessing how recent the data is and checking if the recording time prevents time-delay-related misrepresentation.
Data granularity	Variables and values are defined at the lowest (or as low as required) level of detail for the intended use.	Recording lab results to the appropriate decimal place, as required for meaningful interpretation in a medical setting.	Determining if the level of detail is sufficient for the analytics algorithm to learn and discern cases from one another.

Data validity	A match/mismatch between the actual and expected data values of a given variable.	Defining that the variable "Gender" should only include the values: male, female, or unknown.	Verifying that the data values fall within the defined acceptable values or value ranges for each data element.
Data relevancy	All the variables in the data set are pertinent to the study being conducted.	Including only the most influential financial metrics for predicting stock price, while excluding completely irrelevant administrative codes.	Assessing the degree of relation (a spectrum from least to most relevant) of each variable to the specific problem.

3. Taxonomy of Data (Figure 2.2)

Data can be classified into two major categories: **Structured** and **Unstructured or Semistructured**.

A Simple Taxonomy of Data

[Image of the Simple Taxonomy of Data: A tree diagram starts with "Data in Analytics" at the top. It branches to "Structured Data" and "Unstructured or Semistructured Data".

1. Structured Data branches to Categorical and Numerical.

   * Categorical branches to Nominal and Ordinal.

   * Numerical branches to Interval and Ratio.

2. Unstructured or Semistructured Data branches to Textual, Multimedia, and XML/JSON.

   * Multimedia branches to Image, Audio, and Video.]

Data Type Descriptions and Examples

Category	Type	Description	Example(s)
Primary	**Structured Data**	Data organized for computers to process, typically highly organized (tabular).	Tabular data in an Excel spreadsheet, a relational database table.
Primary	**Unstructured or Semistructured Data**	Composed of textual, imagery, voice, and Web content; created for humans and not readily understandable by computers.	Text documents (emails, social media posts), Images, Audio files, Video files, XML/JSON documents.
Structured	**Categorical (Discrete)**	Represents labels of multiple classes used to divide a variable into specific groups (finite number of values, no continuum).	Race, Educational Level, Age Group.
Categorical	**Nominal**	Measurements of simple **codes assigned as labels**, which are **not measurements** and have no implied order.	Marital Status (1=Single, 2=Married, 3=Divorced), Hair Color (Brown, Green, Blue).

Categorical	**Ordinal**	Codes assigned as labels that also represent **rank order** among them.	Credit Score (1=Low, 2=Medium, 3=High), Educational Level (High School, College, Graduate School).
Structured	**Numerical (Continuous)**	Represents numeric values of specific variables (scalable measurements, allows fractional values).	Age, Number of Children, Total Household Income, Temperature.
Numerical	**Interval**	Variables measured on interval scales; the unit of measurement is fixed, but **there is no absolute zero value**.	Temperature on the Celsius scale (0°C is arbitrary, not "no heat").
Numerical	**Ratio**	Variables with a **nonarbitrary zero value** (measurement is the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind).	Mass, Length, Time, Kelvin Temperature (Absolute Zero).

4. Application Case 2.1: Medical Device Company Ensures Product Quality While Saving Money

Summary of the Case Study

Instrumentation Laboratory (IL), a leading medical device company, was **data-rich but analysis-poor**. Their main **problem** was enabling R&D scientists and engineers to easily access and analyze the massive amounts of test data (millions of records stored in SQL Server databases) to ensure product quality and efficiency. Specifically, they needed to monitor sensor performance in cartridges affected by factors like environmental changes and material inconsistencies. Reliance on IT for data access and using one-off analyses in Minitab caused latency and inconsistency.

The **solution** was adopting **Dell Statistica**, an advanced analytics platform that provided the required statistical functionality and trust for the healthcare environment. They chose Statistica because it allowed a business analyst to quickly build **canned analysis configurations** enterprise-wide, eliminating the need for every analyst to run to IT for data or re-create complex tests. It was also easy to deploy and use without extensive programming knowledge.

The **results** included: **saving hundreds of thousands of dollars** by quickly identifying and fixing manufacturing problems (e.g., inaccuracy in chemical formulation) that would otherwise lead to scrapping batches of cards. Other benefits were: ensuring **regulatory compliance** (e.g., performing statistical validations required by the FDA); **ensuring consistency** of results by standardizing analysis configurations; and **saving time** for engineers via proactive alerting.

Review Questions for Application Case 2.1

1. What were the main challenges for the medical device company? Were they market or technology driven? Explain.

- The main challenges were primarily **technology/process-driven** around data access and analysis, stemming from market demands for product quality and efficiency.
    - **Challenge:** Being "data-rich but analysis-poor".
    - **Market-driven need:** The need to maintain high quality and efficiency while dealing with rapidly evolving medical technologies.
    - **Technology/Process Challenge:** Difficulty for R&D staff to easily access and analyze vast amounts of test data stored in SQL databases; using ad-hoc, inconsistent methods (Minitab). This prevented the constant, intelligent monitoring needed to catch issues quickly.

2. What was the proposed solution?

- The proposed solution was to implement the advanced analytics software **Dell Statistica**. The goal was to provide a platform that allowed analysts to easily access data and run **standardized, complex analyses** from a central store, enabling scale and consistency across the enterprise.

3. What were the results? What do you think was the real return on investment (ROI)?

- **Results/Benefits:** Savings of **hundreds of thousands of dollars** by avoiding scrapping faulty batches, ensuring **regulatory compliance** (FDA statistical validations), enforcing

consistency of analysis across the enterprise, and saving engineers valuable time through proactive automated alerting.

- **Real ROI:** The ROI is likely **extremely high**. Preventing just one batch from being scrapped offsets the cost of the system (saving hundreds of thousands of dollars). Furthermore, achieving regulatory compliance is critical for survival in the medical device industry, and the time saved by engineers can be redirected toward innovation, which fuels future revenue.

5. Data Processing Steps (Figure 2.3)

Data preprocessing is a tedious but **crucial** process to convert raw, dirty, misaligned data into a **well-refined form** ready for analytics algorithms.

Phase	Purpose	Example of a Technique Used	Common Risk or Pitfall to Avoid
Data Consolidation	To collect, select, filter, and integrate/merge relevant data from multiple sources into one unified data set.	**SQL Queries** to select and extract data; **Domain Expertise** to filter unnecessary info; **Ontology-driven mapping** to integrate and unify data using synonyms/homonyms.	**Synonyms/Homonyms :** Failing to properly handle different names for the same entity (e.g., "Customer ID" in one system, "Client Key" in another) or the same name for different entities.
Data Cleaning	To identify and deal with missing values, reduce noise, eliminate erroneous/unusual values, and remove duplicate records.	**Imputation** to fill in missing values with an appropriate statistic (mean, median, mode); **Cluster Analysis** or **Simple Statistics** (averages/standard deviations) to identify outliers/noise.	**Improper Imputation:** Filling in missing values with a method (e.g., mean) that biases the distribution, or removing too many records, leading to a loss of valuable data.

Data Transformation	To convert the data into a form appropriate for specific algorithms, mitigating potential biases and preparing for better processing.	**Normalization** (scaling all variables to a range like 0 to 1); **Discretization/Binning** (converting numeric variables like age into categorical ranges like 'low, medium, high'); **Creating Attributes** (deriving a new, more informative variable from existing ones, e.g., Blood-Type Match indicator).	**Bias from Normalization:** Forgetting that one variable's extreme values may still dominate if its original range was vastly different, or misinterpreting the importance of the new scale.
Data Reduction	To decrease the volume and/or number of variables (dimensions) in the data set to a manageable, most relevant subset.	**Dimensional Reduction/Variable Selection** (using Principal Component Analysis or Correlation Analysis); **Random/Stratified Sampling** to reduce the number of cases/records; **Balancing Skewed Data** (oversampling the minority class or undersampling the majority class).	**Biased Sampling:** Using simple random sampling on sorted data, leading to a sample that does not reflect the essence of the complete data set.

6. Application Case 2.2: Improving Student Retention with Data-Driven Analytics

Summary of the Case Study

Student attrition (dropout) is a significant challenge for academic institutions, resulting in financial loss and damage to reputation. Traditional methods were qualitative and survey-based, offering insight but lacking the instruments to accurately **predict** and improve attrition.

The **proposed solution** was a **quantitative research approach** using historical institutional data from student databases to develop predictive models for freshman attrition. The project used 5 years of data ($16,000+$ freshmen records) containing academic, financial, and demographic characteristics.

Crucial **data preprocessing** steps included:

- **Cleaning/Filtering:** Removing international student records due to missing predictive information (e.g., high school GPA, SAT scores).
- **Transformation/Feature Engineering:** Creating new variables like `Earned/Registered` ratio (student resiliency) and `YearsAfterHighSchool` (time taken before college enrollment) to magnify information.
- **Balancing:** Addressing the high skew in the dependent variable (`Second Fall Registered`), where only $\sim20\%$ were 'No' (dropout). The study built and compared models using the original unbalanced data and a manually balanced data set (equal representation of 'Yes' and 'No' samples).

The study employed four classification methods (Neural Networks, Decision Trees, SVM, Logistic Regression) and three ensemble techniques.

- **Results from Unbalanced Data:** Overall accuracy was high (up to 87.23%), but accuracy for the target 'No' class (dropout) was unacceptably **low (less than 50%)**, showing the model was biased toward the majority 'Yes' class.
- **Results from Balanced Data:** Overall accuracy dropped (to $\sim81\%$), but the accuracy for the critical 'No' class **significantly improved** (up to 87.51%). Support Vector Machines and Decision Trees performed the best individually. Ensemble models (Information Fusion) performed slightly better overall ($\sim82.10\%$).

Conclusion: The study proved that data mining methods, given sufficient and properly preprocessed data (especially **balanced data**), can predict freshmen student attrition with approximately 80% accuracy. Sensitivity analysis also identified key predictors like **Earned/Registered ratio**, **Spring Student Loan**, and **Fall GPA**.

7. Concept Map – Descriptive Statistics for Descriptive Analytics

Descriptive Statistics describes the basic characteristics of the data at hand, one variable at a time, summarizing the data to reveal meaningful, easily understandable patterns. It is a critical part of Descriptive Analytics.

Descriptive Statistics for Descriptive Analytics Concept Map

Business Analytics (Descriptive Analytics)

$\downarrow$

Statistics (Characterize the data)

$\downarrow$

Descriptive Statistics (Describing the Sample Data on Hand)

$\downarrow$

Central Tendency Measures (Measures of Location)	**Dispersion Measures (Measures of Spread/Variation)**	**Shape Measures (Distribution Characteristics)**
Purpose: Estimate/describe the central positioning of a variable.	**Purpose:** Estimate/describe the degree of variation in a variable (compactness or lack thereof).	**Purpose:** Characterize the frequency of data points (distribution).
Mean (Arithmetic Average): Sum of values / number of observations. **Best for non-skewed, non-outlier numeric data**.	**Range:** Difference between the largest (Max) and smallest (Min) values.	**Skewness:** Measure of asymmetry (sway) in a unimodal distribution.
Median: The middle value of sorted data. Not affected by outliers/skewness. **Best for skewed or ordinal data**.	**Variance (s^2):** Average of squared deviations from the mean. **Larger value means more spread**.	**Kurtosis:** Measures the peak/tail nature of the distribution (more or less peaked than a normal distribution).

Mode: The most frequently occurring observation. **Most useful for nominal data with a small number of unique values**.	**Standard Deviation (s):** Square root of the variance. **More meaningful measure of spread**.	**(Supporting Visual: Histogram)** Shows the distributional shape of the data.
(Supporting Visual: Box Plot) Shows Mean, Median, and Quartiles.	**Quartiles/Interquartile Range (IQR):** Divide data into four parts. IQR (Q3 - Q1) describes the range of the middle half of scores. **Less affected by outliers/skewness**.	**(Supporting Visual: Box Plot)** The relative position of the median vs. mean and whisker lengths indicate potential skewness.

8. Concept Map – Different Types of Charts and Graphs

The right chart should be selected based on the **analytical purpose**—what question is the user trying to answer.

Chart and Graph Taxonomy Concept Map

What would you like to show in your chart or graph?

$\downarrow$

Relationship (Between variables)	Comparison (Among items or over time)	Distribution (Frequency of data points)	Composition (Parts of a whole)
Purpose: Explore the correlation/association between variables.	**Purpose:** Compare performance across multiple categories or track trends over time.	**Purpose:** Show the frequency/shape of data across ranges/categories.	**Purpose:** Illustrate relative proportions or accumulation of parts.

Two Variables (Continuous): Scatter Plot (existence of trends, concentrations, outliers).	**Over Time (Continuous Data): Line Chart** (track trends/changes over time, e.g., stock price over 5 years).	**Single Variable (Continuous/Numeric): Histogram** (show frequency distribution of a variable, e.g., age distribution of customers).	**Static (Simple Share of Total): Pie Chart** (illustrate relative proportions, best for ≤ 4 categories).
Three Variables (Adding Dimension): Bubble Chart (Scatter plot where size/color of markers add a third dimension).	**Among Items (Categorical Data): Bar Chart** (compare data across multiple categories, e.g., spending by department).	**Single/Two Variables (Position/Spread): Box-and-Whiskers Plot** (Show Median, Quartiles, Min/Max, Outliers; compare distributions across categories).	**Changing Over Time (Relative Difference): Stacked Area/Bar Charts** (Show how a whole changes and how its parts contribute).
Geographic Map (Combined with other charts to display location-based data).	**Specialized Comparison: Gantt Chart** (Project timelines, task duration, overlap).		**Static (Comparison of Components): Tree Map** (Display hierarchical data as nested rectangles; size/color adds dimensions).

	Heat Map (Comparison of continuous values across two categories using color gradient).		

9. Summary – Dashboard Design

A well-designed dashboard is a **visual display of important, consolidated information** on a single screen, allowing information to be digested **at a single glance** and easily explored. Its distinctive feature is its three layers: **Monitoring** (KPIs), **Analysis** (root cause), and **Management** (actions to resolve).

Actionable Guidelines for Effective Dashboards

1. Prioritize and Benchmark KPIs (KPI Selection and Context):

- **Prioritization:** Only present the essential **Key Performance Indicators (KPIs)** that require action.
- **Context/Benchmarking:** Place numbers in context by comparing them to **past values, forecasted values, or industry standards**. This is necessary to determine if a number is "good" or "bad."
- **Visual Hierarchy (Monitoring/Analysis/Management):** Organize the display based on the three layers of information, ensuring a clear path from high-level status to detailed operational data (Monitoring $\rightarrow$ Analysis $\rightarrow$ Management).

2. Ensure Data Quality and Contextual Metadata (Reliability):

- The most important attribute is the **data itself**—ensure it is **reliable, timely, consistent**, and not missing.
- **Metadata:** Wrap the dashboard metrics with **contextual metadata** (e.g., *Where did this data come from? When was the data warehouse last refreshed?*) to answer basic user questions and build trust.
- **Currency/Dynamic View:** Present a **dynamic, real-world view** with timely data refreshes.

3. Optimize Visuals for Cognitive Load (Pre-attentive Processing):

- **Simplicity and Clarity:** The fundamental challenge is to display all required information **clearly and without distraction**. Use the simplest chart/graph that effectively conveys the message (e.g., time series line graphs for trends, bar charts for comparisons).

- **Pre-attentive Cues (Color Consistency):** Use specialized visual objects (e.g., traffic lights, dials, gauges) or **color-coding** to set an **evaluative context** (e.g., Red/Orange/Green for performance). This enables quick assimilation of status (lowering cognitive load).
- **Usability Validation:** Validate the design by a **usability specialist** to ensure it is user-friendly and transparent, as poor usability leads to low adoption.

4. Prioritize Alerts and Facilitate Action:

- **Guided Analytics:** Provide a capability to **guide the "average" user** through the same navigational path as an analytically savvy user.
- **Proactive Alerts:** Prioritize and rank **alerts/exceptions** so that critical behaviors are proactively pushed to users ("information finds the user"), rather than requiring the user to constantly check/poll the data.
- **Business-User Comments:** Enrich the dashboard with a small text box to capture and tag **business-user comments** to put the structured KPIs in context and add perspective.

Common Design Pitfalls to Avoid:

- **Too much information/clutter:** Violating the "single screen" and "at-a-glance" principles by using too many complex or crowded charts.
- **Lack of Context:** Presenting raw numbers or averages without comparing them to a benchmark, target, or past value.
- **Inconsistent Data:** Using data that is not reliable or consistent across different views of the business.
- **Poor Visual Choices:** Using visually attractive but inappropriate charts (e.g., a pie chart for too many categories).