Data Driven KPI (Key Performance Indicators) Insight & Prediction. Case Studies: Deutsche Bahn AG

Hochschule für Technik und Wirtschaft Berlin. Erstgutachter: Prof. Dr. Uwe Christians. Zweitgutachter: Prof. Dr. Beate Bergter.

Cevi Herdian

2019-04-11

Contents

Chapter 1

Introduction

1.1 Acknowledgments

After a fast 8 years of hard work with the study at Berlin University of Applied Sciences (HTW Berlin). From studienkolleg till master's degree. I am now ready to face new challenges in the real world in my country Indonesia. Writing this thesis has been an interesting experience an also my last work at the university. I have faced some difficulties but none that have stopped me to complete this thesis. This thesis would not be possible without help from the amazing people at that I meet in university and works.

This thesis based on my experience as an intern and working student in a different company from 2015 till February 2019. And I took the use case for this thesis in my last works in Deutsche Bahn Company from September 2019 till February 2019.

I would like to announce special thanks to my parents and my wife. And all of my family. I would also like to thank Prof. Dr. Christians and Prof. Dr. Beate Bergter who has inspired me.

Last but not least I would like to thank my friends, Christ, Mathel, Jabr, Ouafaa, and Hourya. They have been a continuous source of inspiration and always been very helpful.

1.2 Abstract

Background: Big data is as high volume and variety of data. It can be brought together and analyzed at high velocity (speed). Big data used to discover patterns and make better decisions. Deutsche Bahn as a multinational company have also data growth from the company activities. The KPI (Key Performance Indicators) helps a Big Data to more understandable in easy visualization. A right KPI should act as a gauge (measuring tools), helping the company understand whether the company is taking the right path toward the strategic goals. I have decided to investigate further how Deutsche Bahn operates with KPI Big Data.

Purpose: The purpose of this thesis is to look at a Big Data Key Performance Indicators used at Deutsche Bahn Headquarters. Investigate how a Big Data KPI check their way of daily monitoring. See if there is room for further improvements and if the findings are applicable in other company (another Deutsche Bahn company or subsidiary).

Method: I decided to do only quantitative-approach when collecting data. The quantitative data was imported from SAP Business Object Web Intelligence database. Because of data privacy at Deutsche Bahn, the data that I used in this thesis isn't data from Deutsche Bahn self, but I tried to get similar way data from other sources. I created the analysis problem-solving from scratch (from the beginning, especially without making use or relying on any earlier work for help).

Conclusions: Deutsche Bahn uses KPI to increase sales, profit and to get useful information from their big data that can be analyzed. It is up to each sub-departments to decide if and how they want to work with the KPI. Deutsche Bahn Headquarters is successful when operating with big data KPI. The teams get motivated by working with a Big Data KPI and feel that they can affect the outcome to at least a sufficient extent. There are not many negative things to say about how Deutsche Bahn Headquarters operates with KPIs but there is room for improvements. We believe that other departments might be inspired by how Deutsche Bahn Headquarters operates with KPI.

Chapter 2

Data is Everywhere

Do you have a Smartphone? Of course, this kind of questions is not relevant in this era. From our laptop, tablet, PC or our smartphone. These device creates more data today. Data is getting bigger. The world has an astronomical amount of data, an amount that grows larger and larger each day. This Big Data has changed the way the world interacts, uncovered breakthroughs in fast all of our life, from e-commerce, medicine, genetic, financial, laws, and crime, etc.

The principle of a Big Data works that the more the company has data, the more accurately the company get new **insights** and make **predictions** about what will happen in the future. Also revealed new ways to understand trends in business. By comparing more and more data and creating relationships that were previously hidden enable the company to learn and make smarter decisions on targeting business values such as sales, production, marketing or financial situations.

Based on `forbes.com [1]` only **53%** have big data strategies around the world. The top three use case is **retail**, **social analysis**, and **predictive maintenance**. Key points include the following:

- Reporting, dashboards, advanced visualization, self-service business intelligence, and data warehousing are the top five technologies (Figure 2.1).
- Only 53% of companies are using big data analytics today, it grew 17% in 2015. Telecommunications and Financial Services Industries are the fastest adoptions (Figure 2.2).
- The most important big data analytics use case in 2017 is data warehouse optimization. The seconds are customer/social analysis and predictive maintenance (Figure 2.3).
- Big data analytics use cases vary significantly by industry. Data warehouse optimization dominated Financial Services and Healthcare Industry. Customer / Social Analysis is the leading use case in technology-based companies (Figure 2.4).
- The first big data infrastructure is Spark. And The other two most popular software for big data infrastructure is MapReduce and Yarn (Figure 2.5).
- Spark SQL, Hive, HDFS, and Amazon S3 are top three the big data access methods by respondents (Figure 2.6).
- Machine learning with Spark Machine Learning Library (MLib) got more industry support and investment plans. The growth projected by 60% in the next 12 months (Figure 2.7).

The Big Four Consultant (Deloitte, Ernst & Young, KPMG and PricewaterhouseCoopers) are the four biggest professional services networks in the world. They did also so many research about this theme. One of the interesting information is in the paper `"Gut & gigabytes" from PricewaterhouseCoopers (PwC) [2]`.

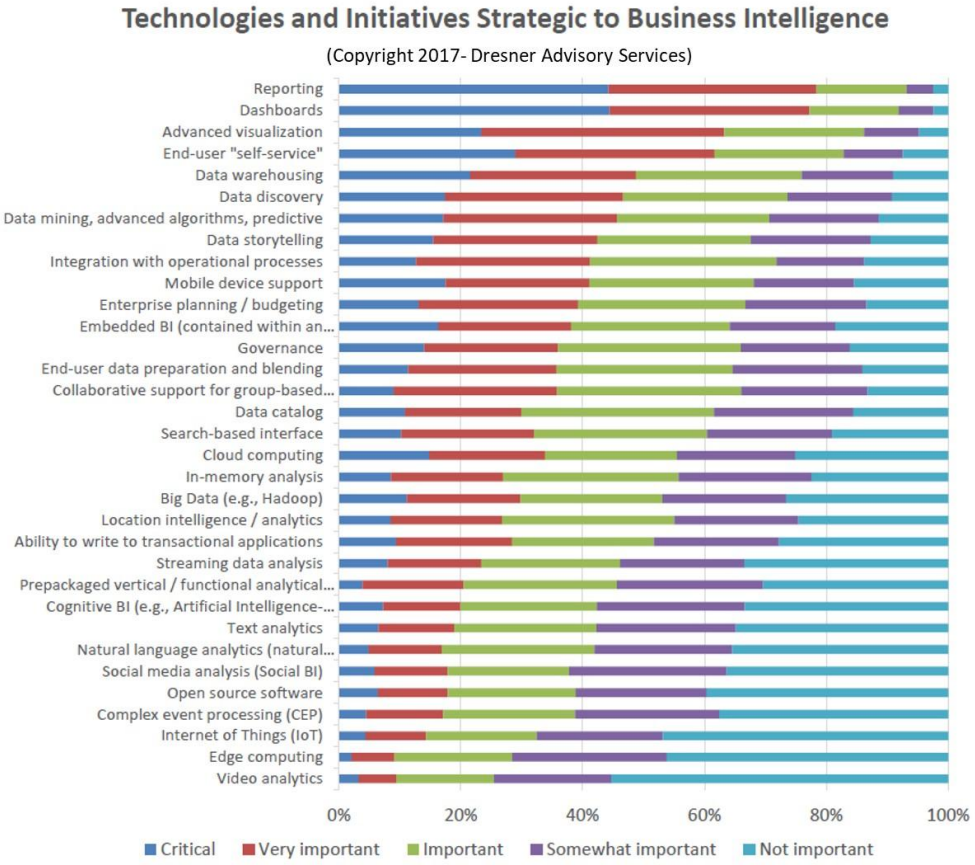
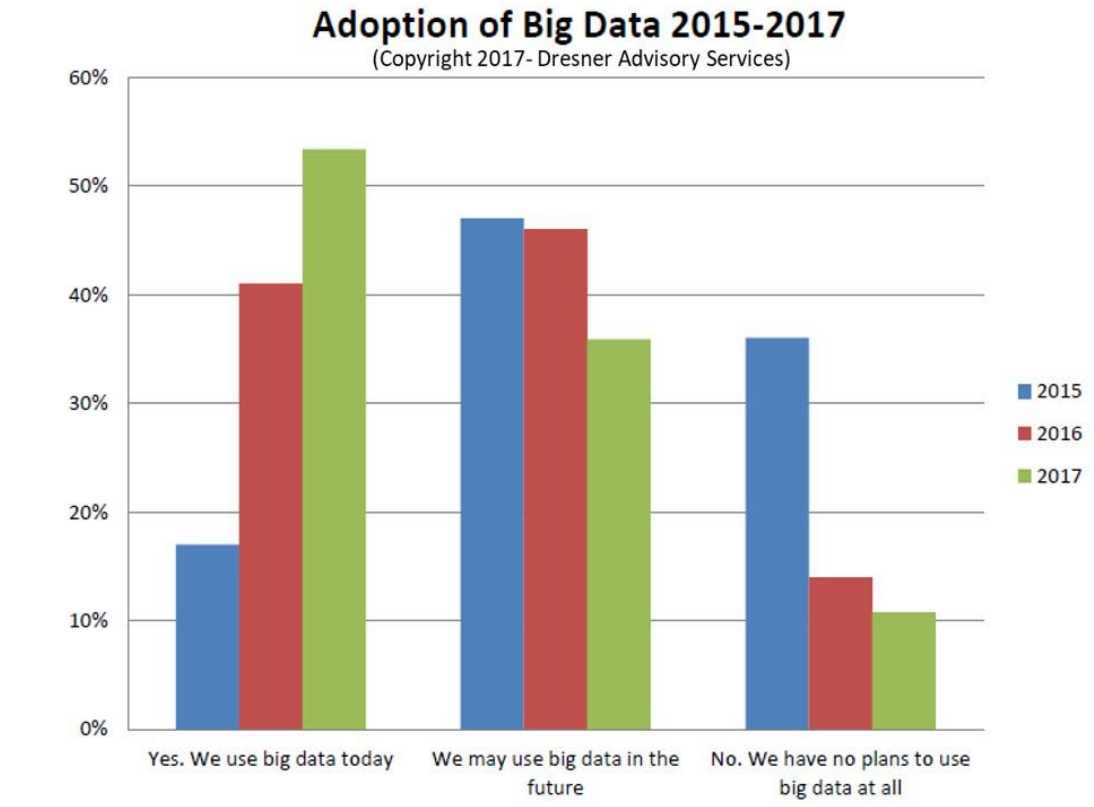

Technologies and Initiatives Strategic to Business Intelligence
(Copyright 2017- Dresner Advisory Services)
Reporting
Dashboards
Advanced visualization
End-user "self-service"
Data warehousing
Data discovery
Data mining, advanced algorithms, predictive
Data storytelling
Integration with operational processes
Mobile device support
Enterprise planning / budgeting
Embedded BI (contained within an...
Governance
End-user data preparation and blending
Collaborative support for group-based...
Data catalog
Search-based interface
Cloud computing
In-memory analysis
Big Data (e.g., Hadoop)
Location intelligence / analytics
Ability to write to transactional applications
Streaming data analysis
Prepackaged vertical / functional analytical...
Cognitive BI (e.g., Artificial Intelligence-...
Text analytics
Natural language analytics (natural...
Social media analysis (Social BI)
Open source software
Complex event processing (CEP)
Internet of Things (IoT)
Edge computing
Video analytics
0%
20%
40%
60%
80%
100%
Critical
Very important
Important
Somewhat important
Not important

Figure 2.1:


Adoption of Big Data 2015-2017
(Copyright 2017- Dresner Advisory Services)
60%
50%
40%
30%
20%
10%
0%
Yes. We use big data today
We may use big data in the future
No. We have no plans to use big data at all
2015
2016
2017

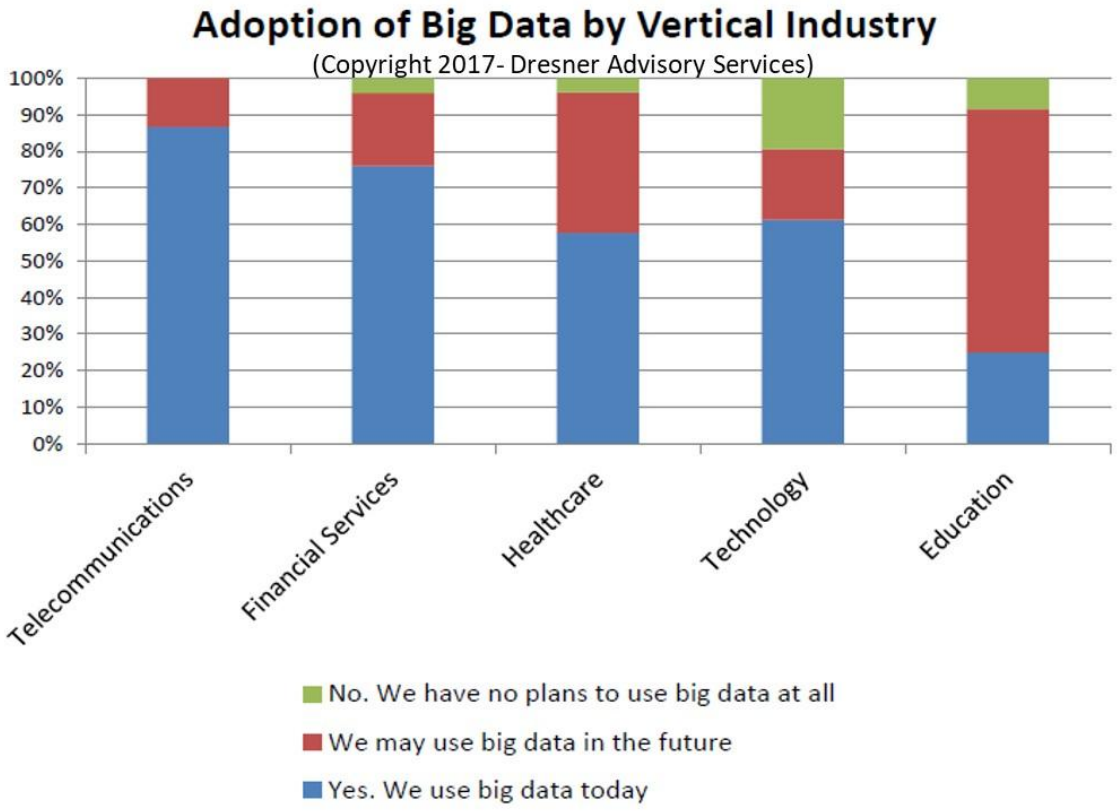

Adoption of Big Data by Vertical Industry
(Copyright 2017- Dresner Advisory Services)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Telecommunications
Financial Services
Healthcare
Technology
Education
No. We have no plans to use big data at all
We may use big data in the future
Yes. We use big data today

Figure 2.2:

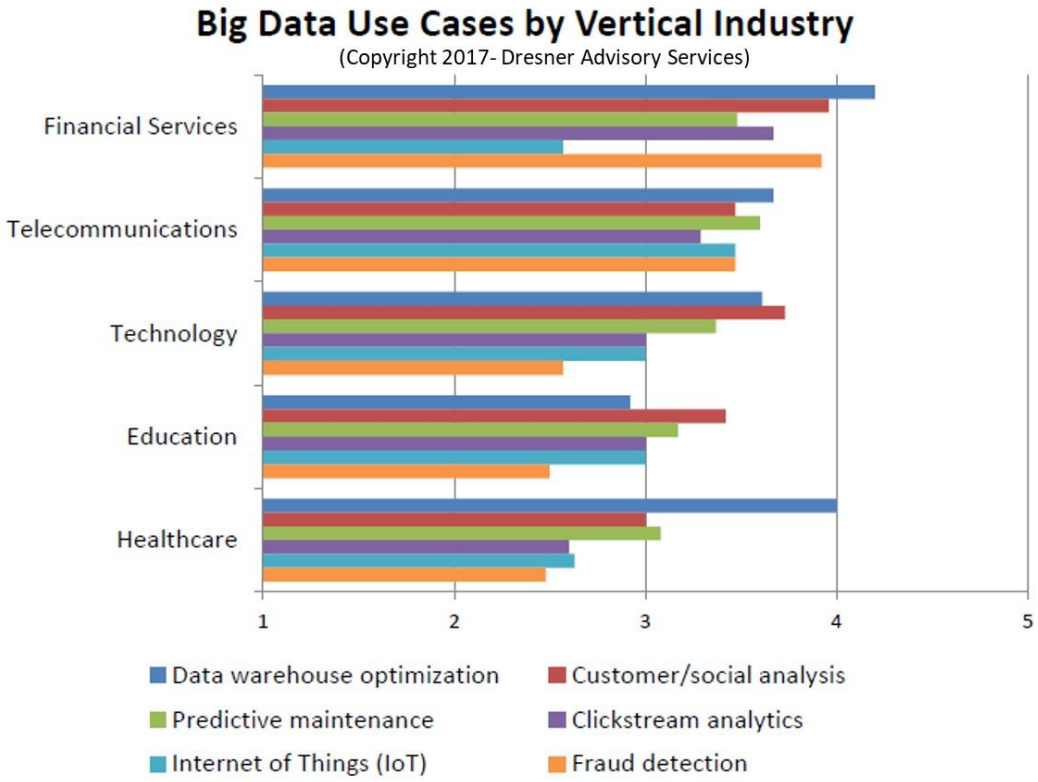

Big Data Use Cases by Vertical Industry
(Copyright 2017- Dresner Advisory Services)
Financial Services
Telecommunications
Technology
Education
Healthcare
1
2
3
4
5
Data warehouse optimization
Customer/social analysis
Predictive maintenance
Clickstream analytics
Internet of Things (IoT)
Fraud detection

Figure 2.3:

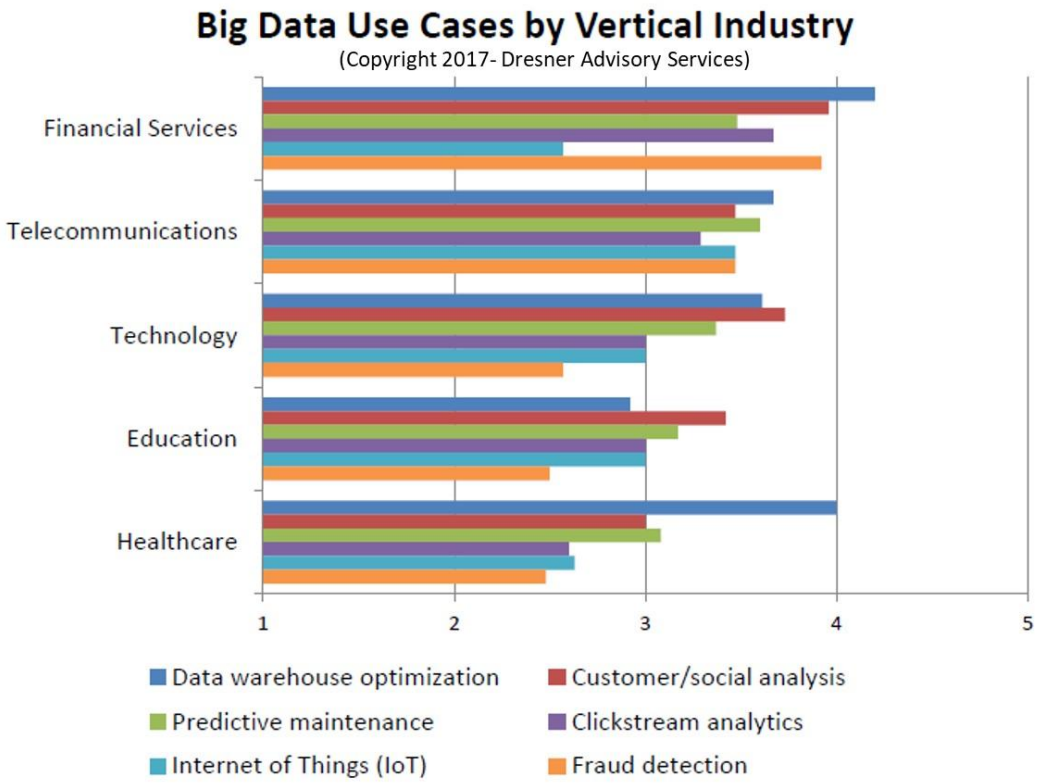

Big Data Use Cases by Vertical Industry
(Copyright 2017- Dresner Advisory Services)
Financial Services
Telecommunications
Technology
Education
Healthcare
1
2
3
4
5
Data warehouse optimization
Customer/social analysis
Predictive maintenance
Clickstream analytics
Internet of Things (IoT)
Fraud detection

Figure 2.4:

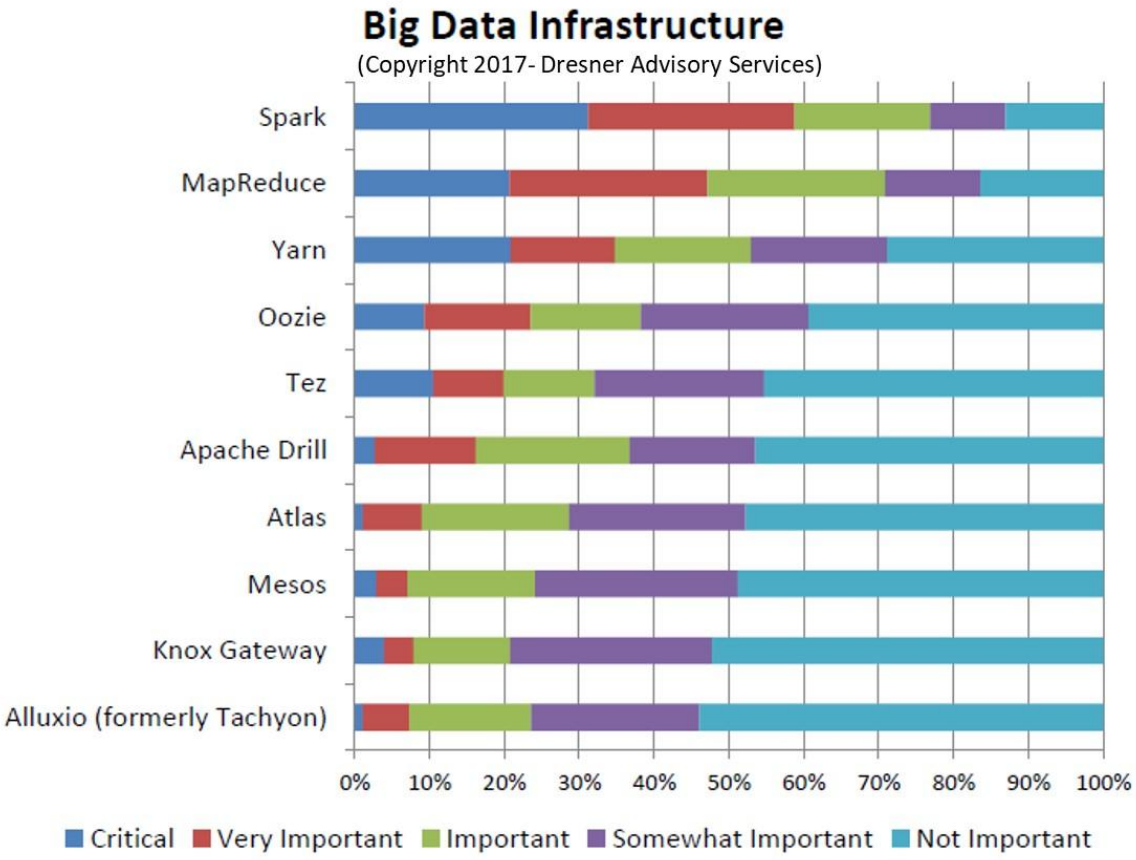

Big Data Infrastructure
(Copyright 2017- Dresner Advisory Services)
Spark
MapReduce
Yarn
Oozie
Tez
Apache Drill
Atlas
Mesos
Knox Gateway
Alluxio (formerly Tachyon)
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Critical
Very Important
Important
Somewhat Important
Not Important

Figure 2.5:

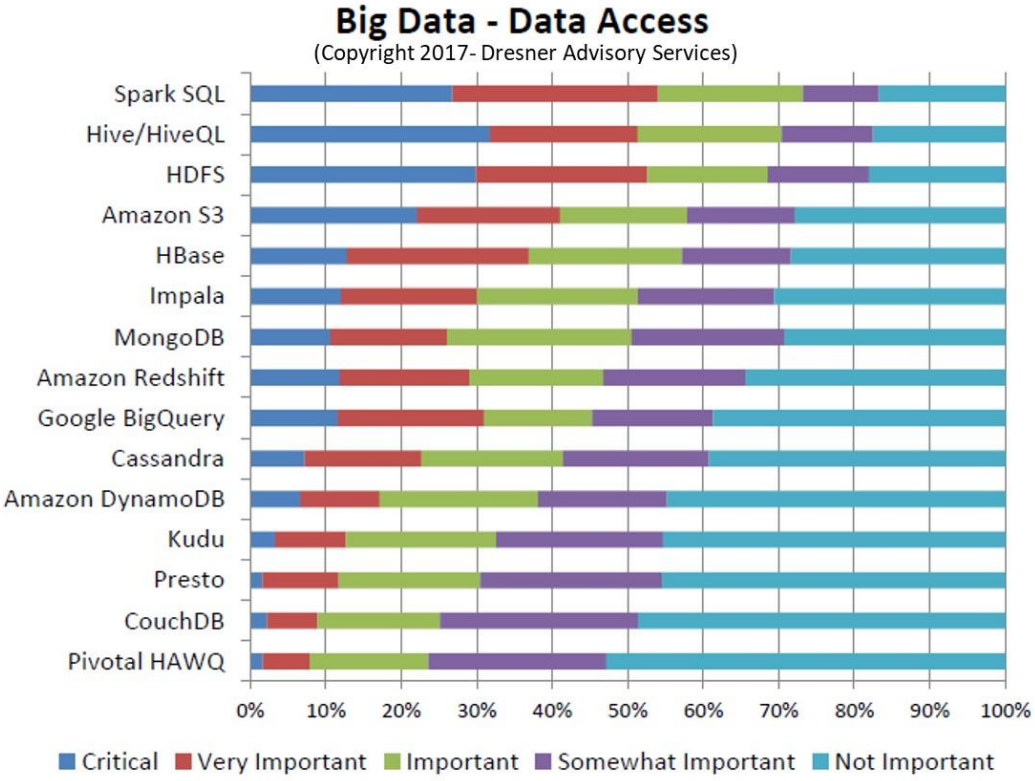

Big Data - Data Access
(Copyright 2017- Dresner Advisory Services)
Spark SQL
Hive/HiveQL
HDFS
Amazon S3
HBase
Impala
MongoDB
Amazon Redshift
Google BigQuery
Cassandra
Amazon DynamoDB
Kudu
Presto
CouchDB
Pivotal HAWQ
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Critical
Very Important
Important
Somewhat Important
Not Important

Figure 2.6:

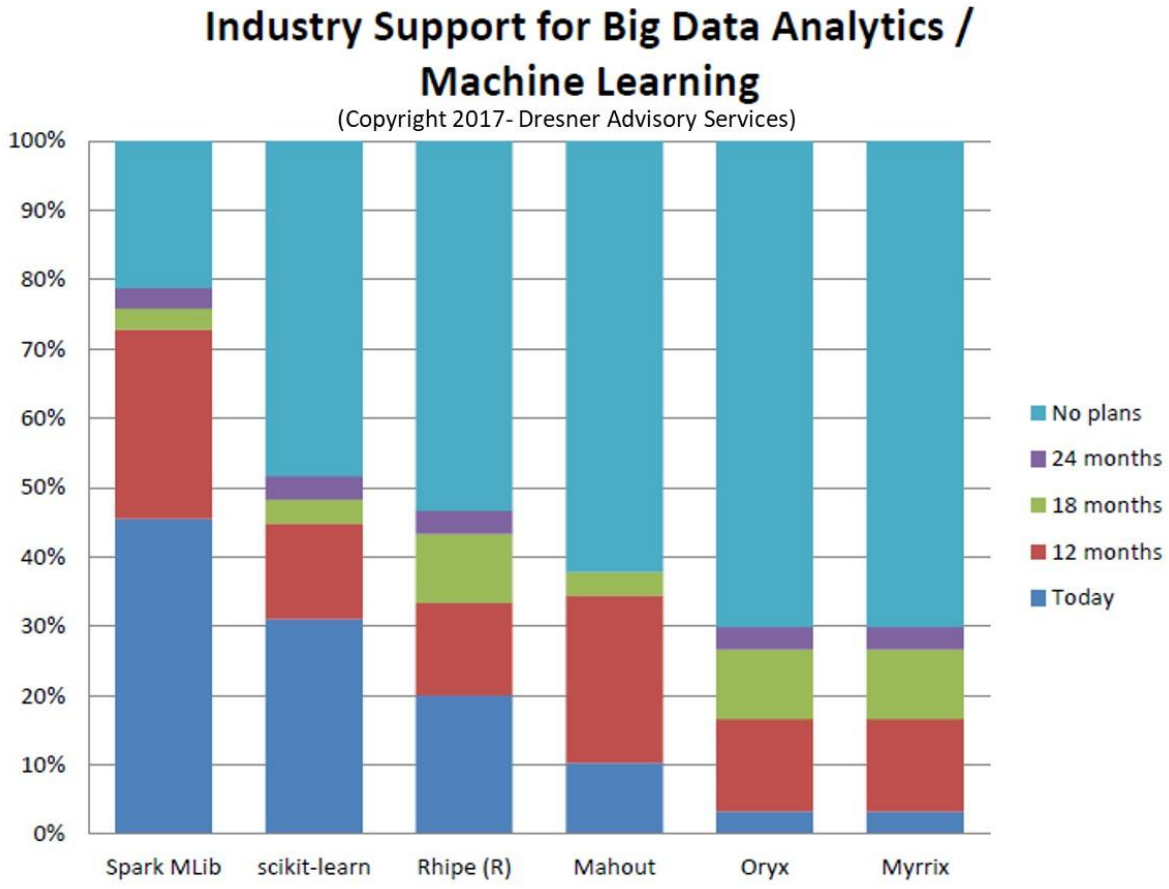

Industry Support for Big Data Analytics / Machine Learning
(Copyright 2017- Dresner Advisory Services)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Spark MLib
scikit-learn
Rhipe (R)
Mahout
Oryx
Myrrix
No plans
24 months
18 months
12 months
Today

Figure 2.7:

The point is that executives rely more on experience and advice than data to make business-defining choices but the data-driven cultures could be impacted more for decision-making.

The summary of this paper:

- Highly data-driven companies are three times more likely to report significant improvement in making big decisions, but only 1 in 3 executives say their organization is highly data-driven.
- More big decisions are made opportunistically than deliberately, and big decisions have a big impact on future profitability; nearly 1 in 3 executives value those decisions at $1 billion+
- Many executives skeptical or frustrated by the practical application of data and analytics for big decisions, especially in emerging markets

And the **five important** issues in this survey are growing the existing business, collaborating with competitors, shrinking the existing business, entering a new industry or starting a new business, and corporate financing.

The last questions of this introduction are how data analytics can improve your businesses? Below are the answered.

1. Make data-driven business decisions:
   - Making evidence-based rather than intuition-based decisions
2. Grow your business – discover new opportunities
   - Quickly identify future markets and the best areas for new investments
   - Boost growth through strategic pricing models and data-driven marketing
3. Create a more efficient and smarter organization
   - Predict and anticipate the impacts of economic, market and regulatory forces on business strategy and results
   - Use automation and advanced statistical software to handle and analyze huge volumes of data
4. Manage risk and regulatory
   - Minimize compliance risks by ensuring the completeness, accuracy, and availability of data sources

2.1 Big Data Definition

As data becomes bigger, manipulating of available data to get insights and make business decisions can be useful. Statistical methods, artificial intelligence, machine learning, and data manipulation are some of the new terms that every business leaders at every level need to become data literate and be able to understand data and analytical concepts.

Before going further, what is actually Big data?. Big data term comes from `John Mashey` in 1990, a computer scientist from Pennsylvania State University. It all starts with the big bang explosion in the amount of data we have created since the rise of the digital era. Side effects of the rise of computers, the Internet and technology that capable of capturing data from all kind of electronic processing.

Big Data has 3 defining properties call 3V. This term introduces by `Gartner analyst Doug Laney` in a 2001 MetaGroup research publication. He published that publication with the title "3D Data Management: Controlling data volume, variety, and velocity". Also, not all data is big data except they have 3 properties. Some other data expert added another 2V, value, and veracity, but the main term is always 3V.

1. Volume: Volume refers to the huge amounts of data generated each time from all the electronic device such as website, social media, cell phones, cars, credit cards, sensors, photographs, video, etc. Volume is the 'V' most associated with the term of big data because volume can be incredible exploding big. Below is the explanation from sisense.com (Figure 2.8)

For Example in the internet world. Facebook is storing 250 billion images As far back as 2016, Facebook had 2.5 trillion posts. Can you imagine for the other online media such as twitter, blog, instagram and our searching on google? It is an extreme amount of data. This link shows the live statistics of some online area such as facebook, twitter., internet users, etc.

http://www.internetlivestats.com/

(Figure 2.9 and Figure 2.10)

Or

http://www.worldometers.info/

(Figure 2.11 & Figure 2.12)

The world information such as population, health, economics, food, water, energy, etc.

2. Variety: It refers to many sources and data types (structured and unstructured). Not only spreadsheet and database but also emails, photos, videos, monitoring devices, audio, etc.

3. velocity: There is four velocity type in big data.

- Batch
- Periodic
- Near real-time
- Real-time

"Firehose" data sources such as social media and e-commerce need quickly analytics. Using real-time alerting, `Wal-Mart` company was able to find a particular sport shoes where it wasn't selling at all. More than realtime but predictive analytics.

The infographics & animations to understand more of the 3V of big data from IBM are below (Figure 2.13).

There has been rapidly growing in the field of Big Data with the benefits of rising the new technology. This availability of big data has led to the use of big data in multiple industries ranging from

- Banking
- Healthcare
- Energy
- Technology
- Consumer
- Manufacturing
- etc

The next section is example use case big data in some of the industries (our case studies: transportation and logistics industries and another industry).

2.2 T & L Industry: Deutsche Bahn

T & L industry is the transport and logistics industry. Deutsche Bahn AG is a German railway company as the second-largest transport company in the world, after the German postal and logistics company Deutsche Post / DHL. Headquartered in Berlin, it is a private joint-stock company (AG), with the Federal Republic of Germany being its single shareholder `(wikipedia.com)`

Based on `Investor Relations site of The Deutsche Bahn [3]`. The Deutsche Bahn Group is divided with some subsidiaries, such as:

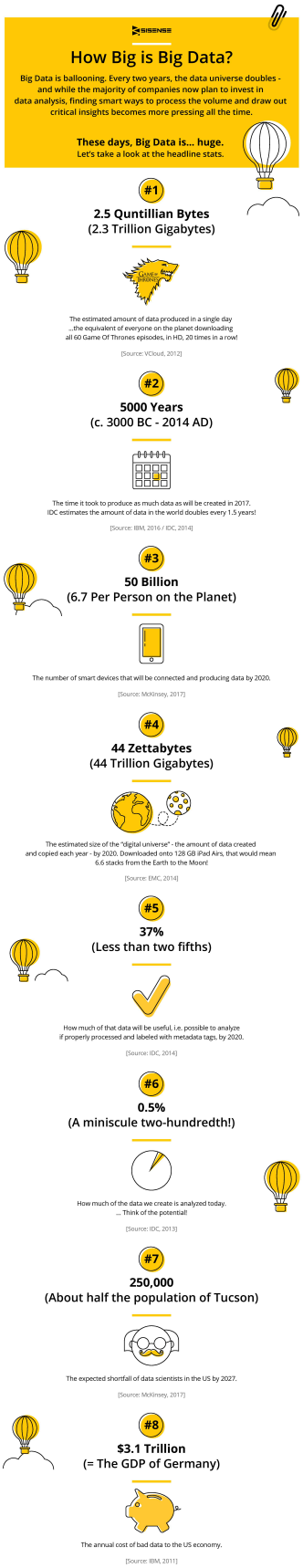

SISENSE
How Big is Big Data?
Big Data is ballooning. Every two years, the data universe doubles - and while the majority of companies now plan to invest in data analysis, finding smart ways to process the volume and draw out critical insights becomes more pressing all the time.
These days, Big Data is... huge.
Let's take a look at the headline stats.
#1
2.5 Quntillian Bytes
(2.3 Trillion Gigabytes)
The estimated amount of data produced in a single day
...the equivalent of everyone on the planet downloading
all 60 Game Of Thrones episodes, in HD, 20 times in a row!
[Source: VCloud, 2012]
#2
5000 Years
(c. 3000 BC - 2014 AD)
The time it took to produce as much data as will be created in 2017.
IDC estimates the amount of data in the world doubles every 1.5 years!
[Source: IBM, 2016 / IDC, 2014]
#3
50 Billion
(6.7 Per Person on the Planet)
The number of smart devices that will be connected and producing data by 2020.
[Source: McKinsey, 2017]
#4
44 Zettabytes
(44 Trillion Gigabytes)
The estimated size of the "digital universe" - the amount of data created
and copied each year - by 2020. Downloaded onto 128 GB iPad Airs, that would mean
6.6 stacks from the Earth to the Moon!
[Source: EMC, 2014]
#5
37%
(Less than two fifths)
How much of that data will be useful, i.e. possible to analyze
if properly processed and labeled with metadata tags, by 2020.
[Source: IDC, 2014]
#6
0.5%
(A miniscule two-hundredth!)
How much of the data we create is analyzed today.
... Think of the potential!
[Source: IDC, 2013]
#7
250,000
(About half the population of Tucson)
The expected shortfall of data scientists in the US by 2027.
[Source: McKinsey, 2017]
#8
$3.1 Trillion
(= The GDP of Germany)
The annual cost of bad data to the US economy.
[Source: IBM, 2011]

Figure 2.8:

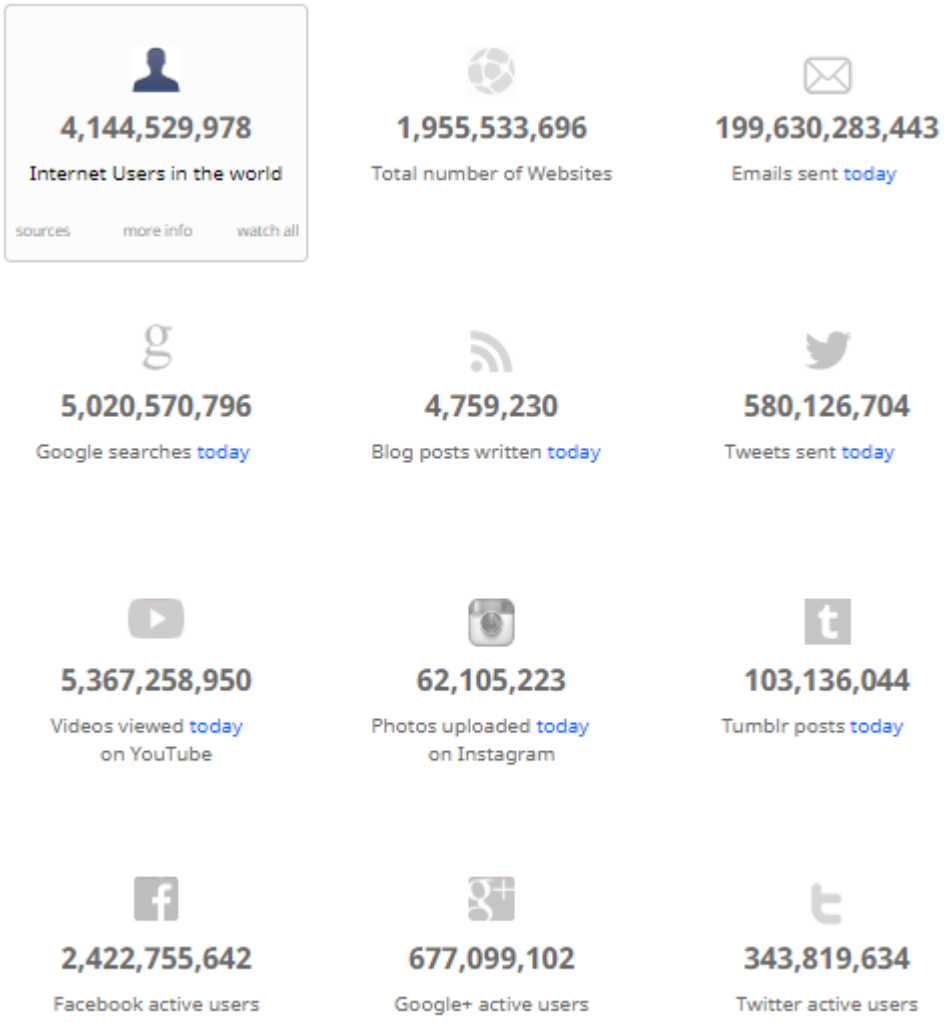

4,144,529,978
Internet Users in the world
sources
more info
watch all
1,955,533,696
Total number of Websites
199,630,283,443
Emails sent today
5,020,570,796
Google searches today
4,759,230
Blog posts written today
580,126,704
Tweets sent today
5,367,258,950
Videos viewed today on YouTube
62,105,223
Photos uploaded today on Instagram
103,136,044
Tumblr posts today
2,422,755,642
Facebook active users
677,099,102
Google+ active users
343,819,634
Twitter active users

Figure 2.9:

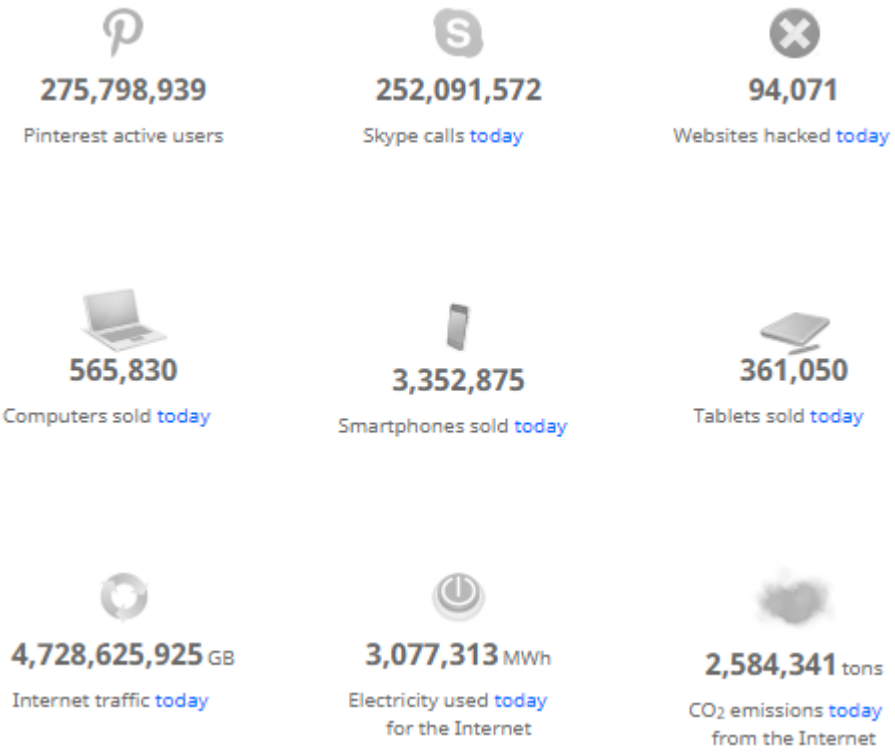


Figure 2.10:

Figure 2.11:


HEALTH
1,380,025 Communicable disease deaths this year [+]
808,032 Deaths of children under 5 this year [+]
4,465,971 Abortions this year [+]
32,858 Deaths of mothers during birth this year [+]
40,625,770 HIV/AIDS infected people [+]
178,706 Deaths caused by HIV/AIDS this year [+]
873,077 Deaths caused by cancer this year [+]
104,273 Deaths caused by malaria this year [+]
12,221,185,146 Cigarettes smoked today [+]
531,424 Deaths caused by smoking this year [+]
265,879 Deaths caused by alcohol this year [+]
113,996 Suicides this year [+]
$ 42,527,306,222 Money spent on illegal drugs this year [+]
143,501 Road traffic accident fatalities this year [+]

Figure 2.12:

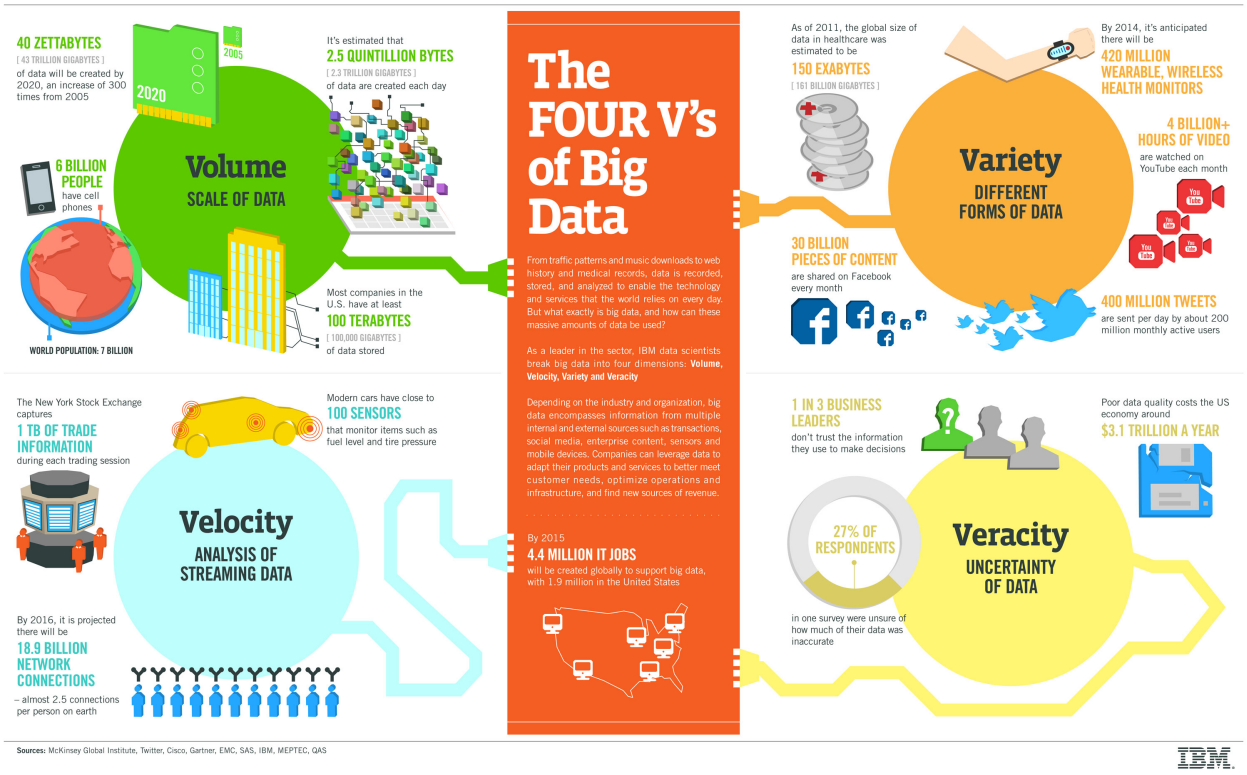


Figure 2.13:

- Personenverkehr
- Arriva
- DB Fernverkehr
- DB Regio
- DB Netze
- DB Engineering & Consulting
- Logistics
- etc

(See Figure 2.14)

Deutsche Bahn Group (DB Group) Key facts:

- Operating globally in more than 130 countries.
- More than 310,000 employees
- Almost 40% employed outside Germany.
- More than 12.5 million people each day on the trains and buses
- Transports over 270 million ton of goods per year by rail,
- Transports over 100 million shipments by road
- About 33,000 km rail network

See Figure 2.15 and 2.16.

Retrived from `Investor Relations site of The Deutsche Bahn [3]`. There are four key success factors in the development of DB Group, which are a central component of DB Group's business model:

1. Entrepreneurial approach to business: in the course of the German rail reform DB Group has established itself as a commercial enterprise. Particularly worth mentioning in this context are the establishment of a modern and efficient organization and a value-based management approach with

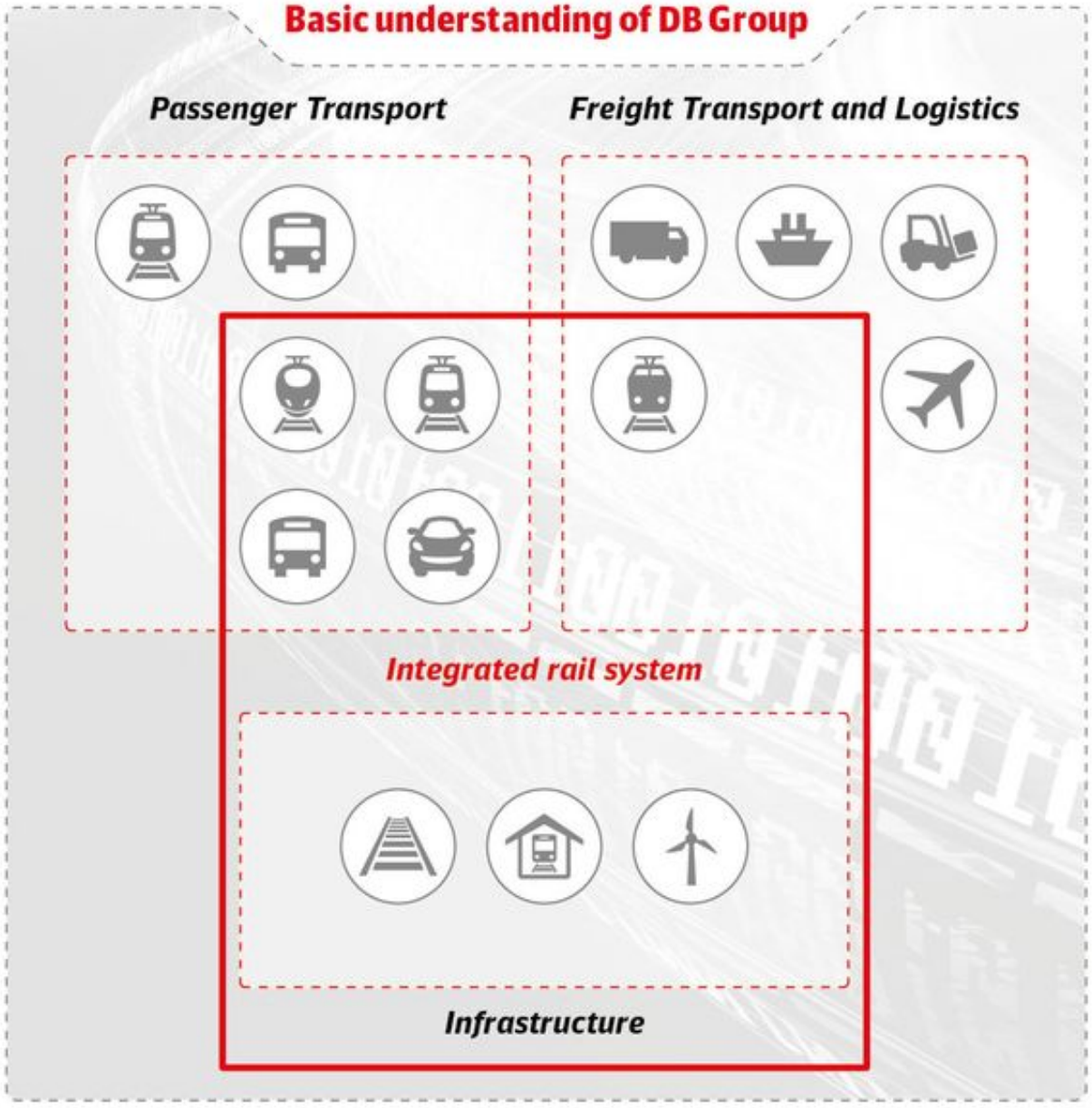

Basic understanding of DB Group
Passenger Transport
Freight Transport and Logistics
Integrated rail system
Infrastructure

Figure 2.14:

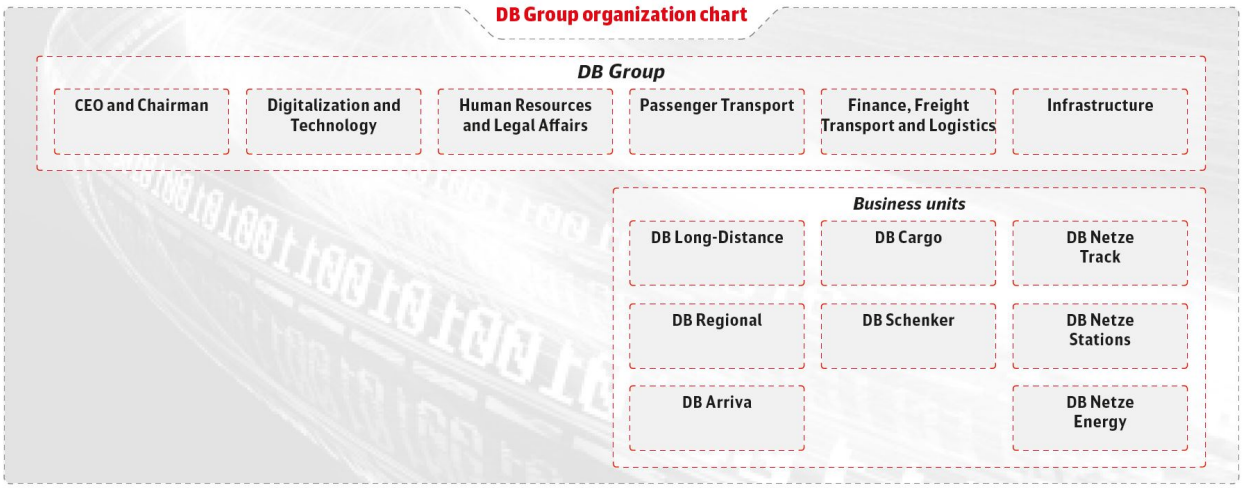


Figure 2.15:

capital market viability as a target.

2. Integrated Group: as a system integrator in Germany, DB Group optimizes the integrated rail system. In doing so, it serves as an important driving force for technological innovation. The integrated Group structure enables us to achieve positive synergies and align our infrastructure to support efficiency, market orientation, and profitability.

3. International position: due to our focus on Europe in passenger transport as well as our European and global orientation in the areas of freight transport and logistics activities, DB Group has an excellent position in the relevant markets. As a result, we are responding to the increasing demand for cross-border solutions. At the same time, we are best positioned to take advantage of growth opportunities.

4. Cross-modal transport solutions: we offer our customers door-to-door mobility and logistics solutions from a single source. We use digital technologies to intelligently link various modes of transport in an economical and environmentally friendly way. In addition, we offer complementary products and services in the freight transport and logistics market.

One of the subsidiary company that focusing on big data analytics to achieve that key success factors is DB Systel Gmbh (https://www.dbsystel.de/dbsystel). `Based on DB Systel Magazine [4]`, the key point of DB Systel are:

Big data

If the data can no longer be stored by conventional databases, it is known as big data. We need mathematical methods, statistical, machine learning, and algorithms to get insight from the big data.

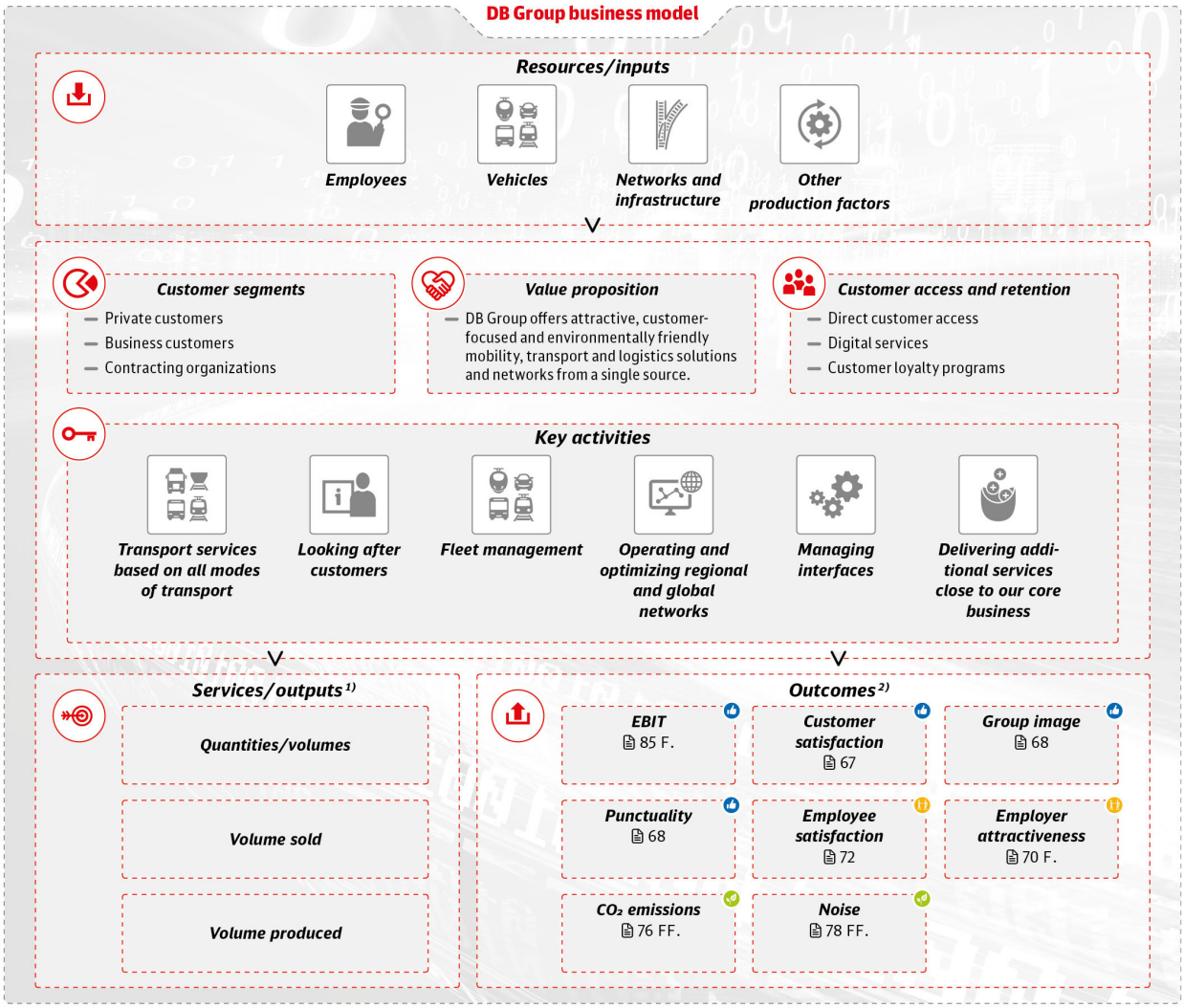
Business intelligence / business analytics

The process to collected, analysed, and visualized data, either bar chart, pie chart or another form. The KPI (Key Performance Indicators) are the main end results from this process. And support decision making in the present and also future

Data mining

The three main concepts of data mining are statistics, mathematics, and algorithms. To detect the hidden patterns and connections in big data we need some methods, such as clustering (formation of groups of similar data), regression analysis (what depends on which other factors?) and association (if one thing happens, another does too)

Data lake



1) Key products and services.

2) Internal and external consequences and profit from business activities and outputs along the entire value-added chain.

Figure 2.16:

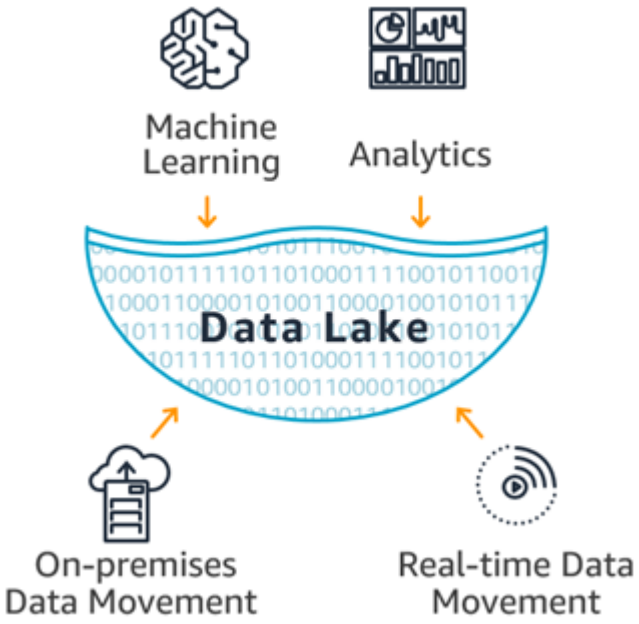


Figure 2.17:

Every day, individual generates around 650 megabytes of data. There is a lake of data. Creating the data lake is a logical way of helping company to get flexible access to all of the data sources. A data lake is a central location where every department can store external and internal data (structured and unstructured data at any scale). (See Figure 2.17)

Data Lakes compared to Data Warehouses. (See Figure 2.18)

Smart data

While big data and the data lake consist of structured data and unstructured data, smart data have a specific purpose using certain algorithms and other tools. Smart data developed and improving business processes and decision-making (See Figure 2.19). The smart data consist also the analytics edge, such as:

- Descriptive analytics
- Predictive analytics

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

Figure 2.18:

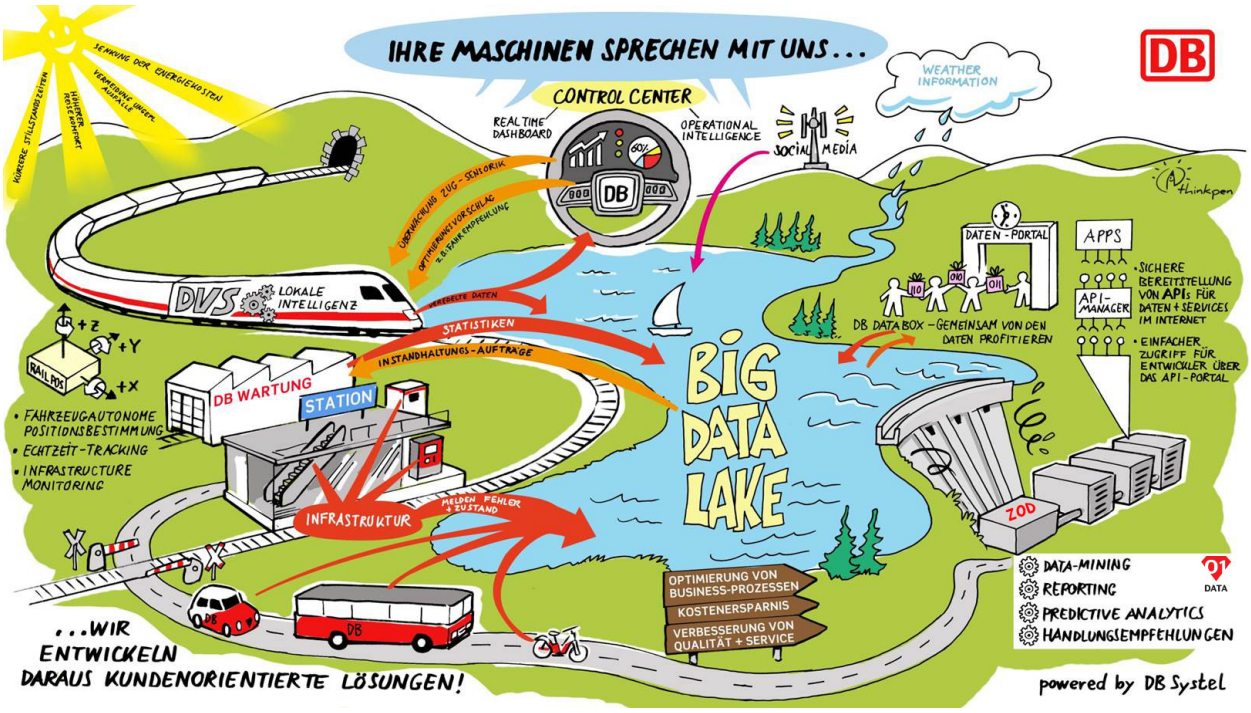


Figure 2.19:

- Prescriptive analytics

The description of these three analytics edges you can find in the next chapter `3. Big Data Analytics`

Like most other industries, transportation and logistics (T&L) are currently confronting immense change, this brings both risk and opportunity. This risk and opportunity have new technology, new market, new customers expectation, and a new business model. Based on `PwC (PricewaterhouseCoopers) research Transport and Logistics Trend Book 2019 [5]`, there are four disruption and uncertainty and five forces transforming in the era of big data for T & L industry (Transportation & Logistic).

Four areas of disruption and uncertainty:

1. How to change customer expectations: (See Figure 2.20)
2. Technological used: (See Figure 2.21)
3. New entrants to the industry
4. Redefining collaboration

Five Forces Transforming Transport & Logistics:

As the complexity of modern transport and logistics grows, it is increasingly difficult to understand what to focus on, so PWC has identified five key forces transforming the T & L segment:

1. Digitalization
2. Shifts in international trade
3. Software-driven process changes
4. Changes in markets domestic commerce
5. Machine-driven process changes

(See Figure 2.22)

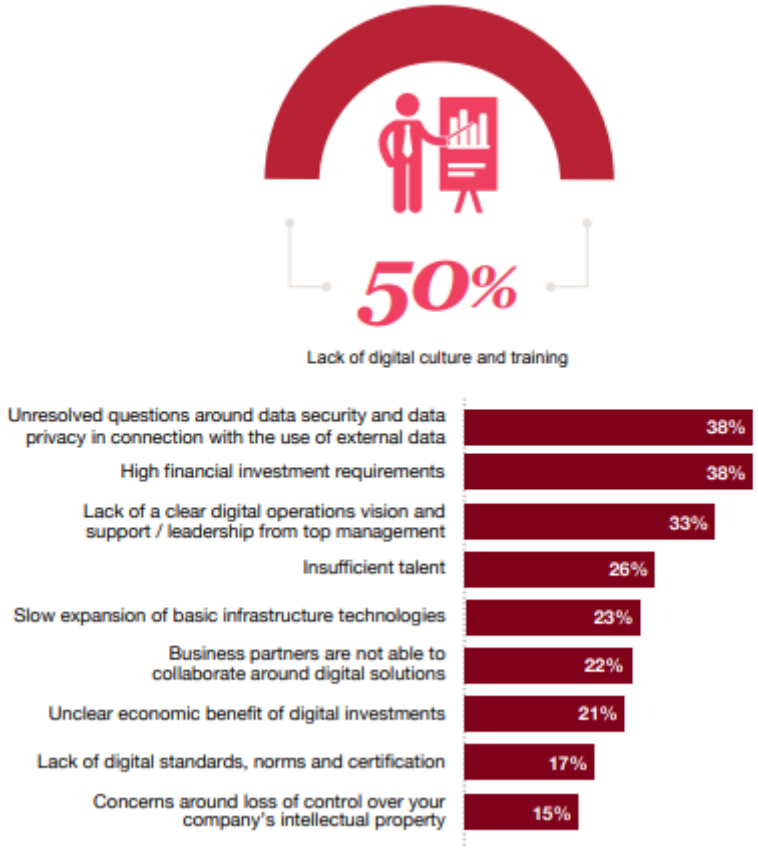

50%
Lack of digital culture and training
Unresolved questions around data security and data privacy in connection with the use of external data
38%
High financial investment requirements
38%
Lack of a clear digital operations vision and support / leadership from top management
33%
Insufficient talent
26%
Slow expansion of basic infrastructure technologies
23%
Business partners are not able to collaborate around digital solutions
22%
Unclear economic benefit of digital investments
21%
Lack of digital standards, norms and certification
17%
Concerns around loss of control over your company's intellectual property
15%

Figure 2.20:

The technology[10]	The impact	The uncertainties
Physical Internet (based on the IoT)	• Improved supply chain transparency, safety and efficiency • Improved environmental sustainability (more efficient resource planning)	• Social expectations around data privacy and security may change • Regulation around data security and privacy may increase or be enforced more stringently • The sector's willingness and ability to invest in collaboration • Whether international bodies will drive standardisation
IT standards	• Enabling collaboration horizontally • More efficiency and transparency	• Companies' willingness to adopt is uncertain due to data security concerns
Data analytics	• Improvements in customer experience and operational efficiency in operations • Greater inventory visibility and management • Improved 'predictive maintenance'	• Rate of development of data processing capacity is unclear • Question marks around data security • Social expectations around data privacy and security may change • Regulation of data security and privacy may increase or be enforced more stringently
Cloud	• Enabling new platform-based business models and increasing efficiency	• Development of costs unclear (once a certain scale is reached physical data centres still tend to be cheaper) • Uncertainties around data security
Blockchain	• Enhanced supply chain security (reduction of fraud) • Reduction in bottlenecks (certification by 3rd parties) • Reduction of errors (no more paper-based documentation) • Increased efficiency	• Rate of adoption uncertain • Unclear whether one or two dominant solutions will emerge or multiple competing solutions
Robotics & automation	• Reduction in human workforce and increased efficiency in delivery and warehousing (including sorting and distribution centres) • Lower costs	• Speed of technology development unclear
Autonomous vehicles	• Reduction in human workforce • Increased efficiency in delivery processes	• Regulatory environments not currently in place in most countries • Liability issues not yet clear • Ethical questions remain especially in relation to emergency situations
UAVs / Drones	• Increased cost efficiency (use cases: inventory, surveillance, delivery) • Workforce reduction	• Regulation in most countries not sufficient for commercial use in public areas like delivery • Safety and privacy concerns may hamper market acceptance
3-d printing	• Lower transportation demand • Transported goods would mostly be raw materials	• Speed, scale, and scope of uptake by customer industries still unclear

Figure 2.21:

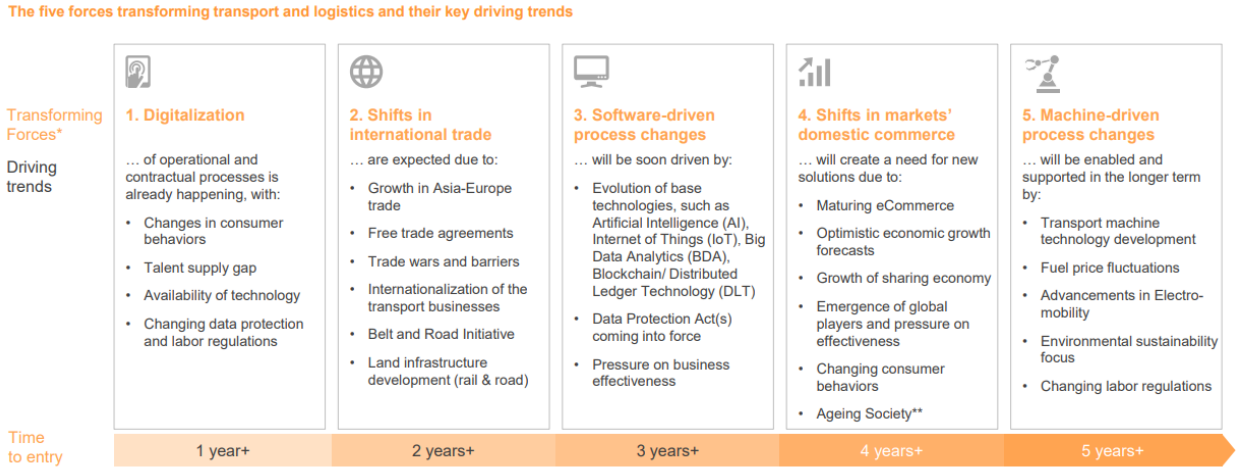


Figure 2.22:

2.3 Another industry

2.3.1 Agriculture

Agriculture businesses are using big data to improve productivity and yields and improve forecasting to better optimize supply chains. As agribusinesses growth and more diverse, the growing of data that must be managed is also becoming more huge and complex. The data comes from not only social media outlets and supplier network channels but also agricultural devices sensor and machine equipment. Today these data sources come from:

- Traditional enterprise data from operational systems
- Farm field sensor data (e.g. temperature, humidity, rainfall, sunlight)
- Farm equipment sensor data (from tractors, plows, and harvesters)
- Harvested goods and livestock delivery vehicles (from farms to processing facilities) sensor data
- Commodities trade data
- Financial data
- Weather data
- Animal and plant genomics research data
- Social media

Big Data can help improve forecasting and efficiency and lead precisely decision making. Big data technologies enable agricultural sectors to analyze a variety of data sources, in turn leading to better outcomes. Agribusinesses have long lists of needed metrics. But which data we need and where the data come from and also what is the outcome or beneficially result. Following is a list of areas where Big Data technologies can impact in agricultural sectors:

- Weather data from weather institution for better understanding time to plants
- Improved forecasting of yields and production data from Farm field sensor data (e.g. temperature, humidity, rainfall, sunlight)
- Better optimized seeds and livestock and new methodologies that improve yields and production with Farm equipment sensor data such as warehouse sensor data
- Faster delivery of goods produced to distribution centers and consumers with street data or vehicle (car) sensor data
- Real-time decisions and alerts based on data from fields and equipment with Farm equipment sensor data or Farm field sensor data
- Integrated production and business performance data for improved decision-making with ERP (Enterprise Resource Planning) system from integrated all sensor devices and sales data.
- Rationalized performance data across multiple geographies with sales data
- Accurate crop predictions with weather and crop data
- Stronger seeds and less hunger with plant data with purpose developing crops that can grow in any environment -Automated agriculture for better production with the different system

(See Figure 2.23)

Sources:

- `Big Data in food and agriculture [6]`
- `Big Data and its impact on agriculture [7]`
- `Big Data in Smart Farming - A review [8]`
- `4 ways big data analytics is disrupting the agriculture industry [9]`

2.3.2 Automotive

The automotive industry faces a complex challenge. One thing that has the opportunity to deliver a solution is analytics. Shifting marketing conditions, volatility, competition, and cost pressure are the problem to change in the automotive industry.

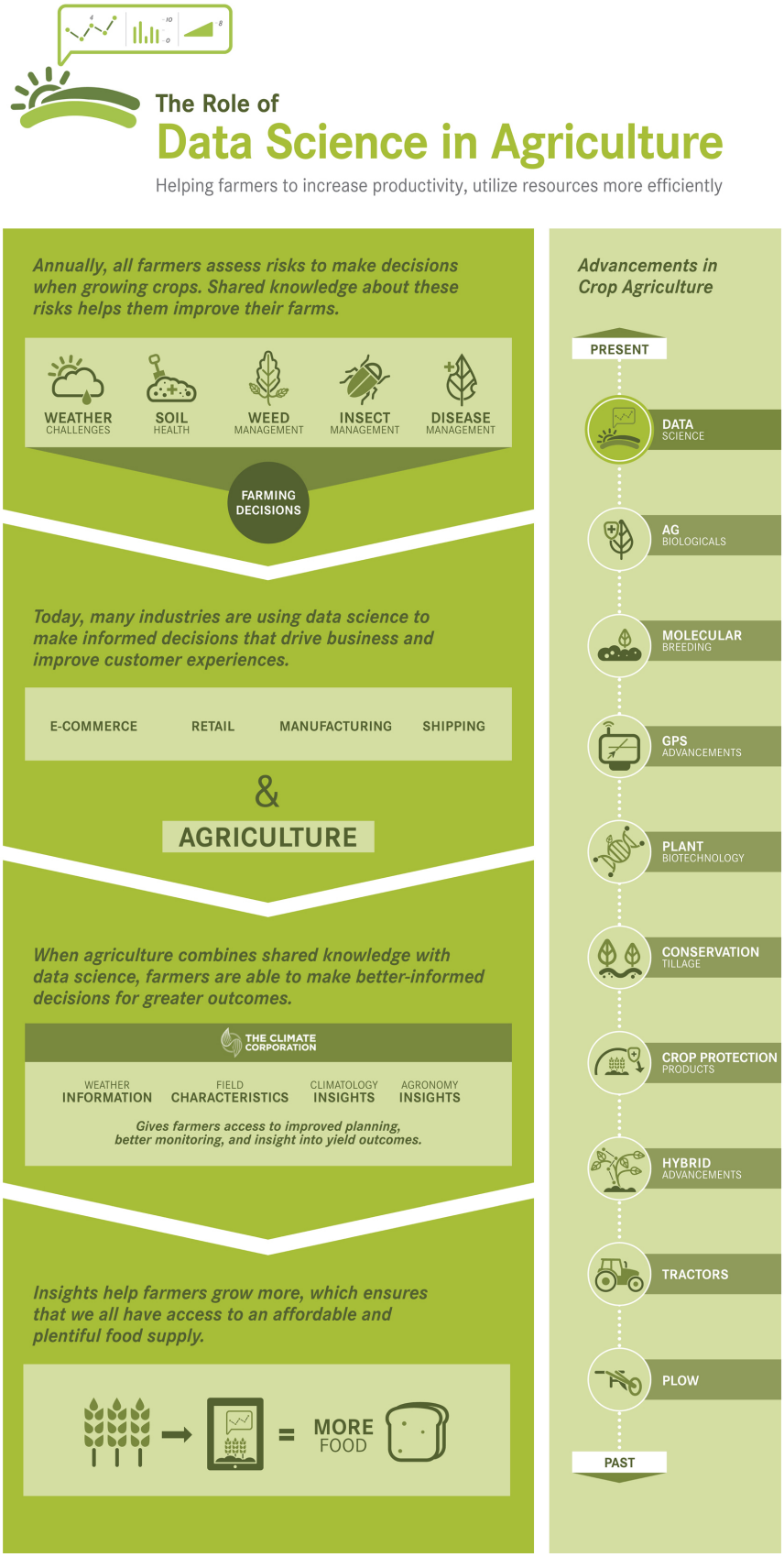

The Role of
Data Science in Agriculture
Helping farmers to increase productivity, utilize resources more efficiently
Annually, all farmers assess risks to make decisions when growing crops. Shared knowledge about these risks helps them improve their farms.
WEATHER CHALLENGES
SOIL HEALTH
WEED MANAGEMENT
INSECT MANAGEMENT
DISEASE MANAGEMENT
FARMING DECISIONS
Today, many industries are using data science to make informed decisions that drive business and improve customer experiences.
E-COMMERCE
RETAIL
MANUFACTURING
SHIPPING
&
AGRICULTURE
When agriculture combines shared knowledge with data science, farmers are able to make better-informed decisions for greater outcomes.
THE CLIMATE CORPORATION
WEATHER INFORMATION
FIELD CHARACTERISTICS
CLIMATOLOGY INSIGHTS
AGRONOMY INSIGHTS
Gives farmers access to improved planning, better monitoring, and insight into yield outcomes.
Insights help farmers grow more, which ensures that we all have access to an affordable and plentiful food supply.
= MORE FOOD
Advancements in Crop Agriculture
PRESENT
DATA SCIENCE
AG BIOLOGICALS
MOLECULAR BREEDING
GPS ADVANCEMENTS
PLANT BIOTECHNOLOGY
CONSERVATION TILLAGE
CROP PROTECTION PRODUCTS
HYBRID ADVANCEMENTS
TRACTORS
PLOW
PAST

Figure 2.23:

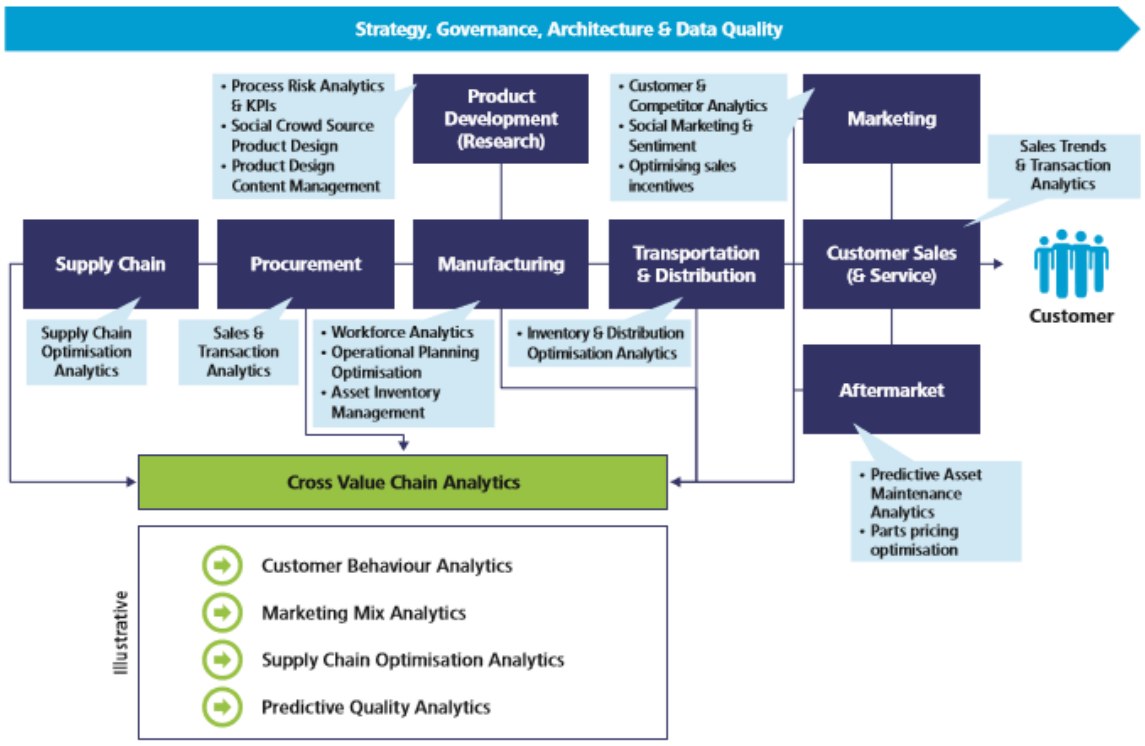


Figure 2.24:

As the automobile is being transformed by big data technologies (everything from sensors to artificial intelligence to big data analysis). The ecosystem is change and the analytics allows the data not only structured but also unstructured data such as videos, sound recording, or texts to be analyzed the results are impressive. (See Figure 2.24)

Below are the key points to deeply undertand the automotive industry:

- **Use the data for actionable customer segments and individualised offers and to boost sales and improve customer retention.**

Customer behavior analytics: Customer need a consistent personalized experience across their access channels. Automakers are having to sell by offering 24/7 connectivity and reliance on social media and the internet as communication tool. There is a huge of data available to support automakers for understanding their customers but automakers have limits and ability to collect, analyze, and act on it.

Automakers need to fully understand customers needs and behaviours to develop a bird view of customer segmentation. For example, the combination of online retailing and physical stores to build brand awareness, attract new target and transformational retailing experience for shopping anytime and anywhere. At least but not last, customer retention needs to be placed high on the strategy, understanding the customer experience with the life cycle and journey not only just during the purchase.

The example of real strategy:

> Inner city mobility: Pay for use customer mobility solution, covering central city locations, offering ease, speed, and flexibility of commute
>
> Urban lifestyle convenience: Taking direct retailing and servicing business to the consumer
>
> Regional retailing: Providing the complete digitized brand experience through existing dealers

- **Applying big data analysis to a mass historical data to identify the impact** of fixed and variable marketing parameters and support auto industry with a more precise and effective approach to quantity and composition of marketing cost.**

The growth in data and information available on customers is allowing automakers and dealers to focus on specific customers. To make this happens, automakers bring together internal and external information sources relating to variable marketing spend to better undertand what‘s working and what isn´t. Besides that, there is some issue that must be overcome, including:

Lack of available data on the transaction for internal and competitor brands.

Less understanding of competitor marketing and strategies

Differences in customer behaviours across the different region and customer segments

Less ability to understand the impact of trends on market and buying behaviours

- **Supply chain data can reveal which links in chain could weaken allowing for proactive and timely countermeasure before a real problem manifests.**

Globalizing operations to take advantage of high-growth markets, driving innovative strategies to optimise manufacturing process. Get it right and gain a competitive advantage and drive growth. Get it wrong and automakers can get difficult scenarios from parts shortages, government security, or lost growth opportunities. Moreover, advanced supply chain analytics can help automakers analyse larger dataset using analytical and mathematical techniques. An example of this is the use of product configuration and web interactions which allow automakers to get early of emerging trends and forecast.

- **Predictive analytics for forecasting efficiency as well as operations and performance.**

The good news is that predictive quality analytics capabilities available today allow quality issues to be detected and taken prevented. This news provided an improved ability to better manage customer satisfaction and cost control concerns.

Source: `Opportunities for analytics in the automotive industry [10]`

2.3.3 Retail

The sphere of the retail industry develops rapidly. Besides that, the retail industry is the top use case of big data. The retailers manage to analyze data from different sources and develop an overview of a customer to learn targeting sales. Besides that, a customer needs to be easily influenced by the strategy developed by another retailer. New sources of data, from machine data (log files) and transaction, to sensor and social media, present opportunities for retailers to get impacted value and competitive advanced. To achieve this is to make the best plans and decisions, understand customer more deeply, uncover hidden trends, and becoming a data-driven organization. (See Figure 2.25)

This example presents top big data analytics use cases in the retail, created for retailers to be aware about trends and tendencies using big data.

1. Customer Behavior Analytics for Retail: Retailers can combine, integrate and analyze all of your external and internal data to generate deep or hidden insights.
2. Customer Journey Analytics: Today’s customers are more connected than before. Using mobile devices, social media and e-commerce, customers can access all of the information in seconds. With big data technologies, retailers can bring data into powerful insights. Retailers will be able to get deeply understanding these questions such as:

- What is s really happening across every step in the customer journey?
- Who are your high-value customers and how they behave?
- How and when is it best to reach them?

3. Personalizing the In-Store Experience With Big Data

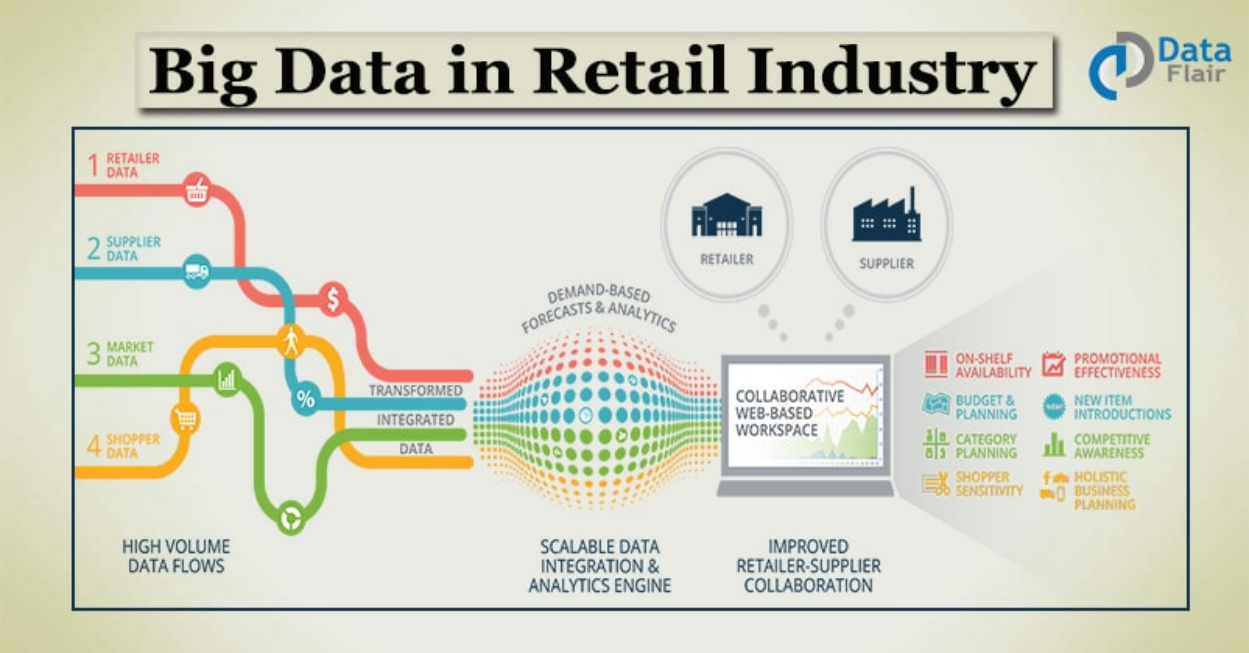


Figure 2.25:

Big data analytics can turn data sources into an advantage for retailers. These data sources can be gathered from:

- Websites
- Point of sales systems
- Mobile apps
- Supply chain systems
- In-store sensors
- Cameras and more

4. Increasing conversion rates through predictive analytics and targeted promotions

To reduce costs and high customer acquisition, retailers need targeting promotions. This requires a 360-degree view of customers situation.

In the past, customer information has been limited to demographic data during sales transactions. But today, customers interact more than they transact, as an example is social media. Using data from internal or external sources, omnichannel retailers can:

- Test and quantify the impact of different promotional tactics on customer behavior and conversion
- Use a customer purchase and browsing history to identify needs and interests and then personalize promotions for customers
- Monitor customer purchasing behavior and social media activity to drive timely offers to customers to incent online purchases with a specific retailer

5. Recommendation engines

(See Figure 2.26)

Recommendation engines are one of the tools for customers behavior prediction. Recommendation engines provided retailers to increase sales and to dictate trends. Recommendation engines depend on the choices by the customers. Usually, these engines use either collaborative filtering or content-based recommendation.

- Collaborative Filtering:

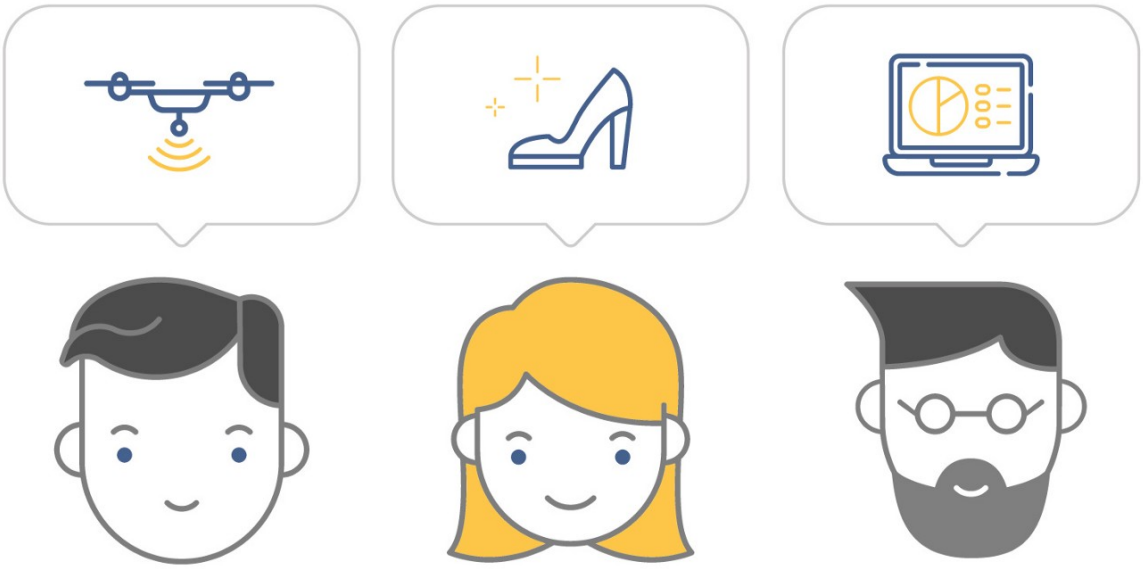


Figure 2.26:

(See Figure 2.27)

This system based on how similar users liked the item. As an example, Cevi and Herdian have similar interests in games. Cevi bought and played "Game of Thrones 4". Herdian has not played this game.

But the system recommends this game because the system has learned that Cevi and Herdian have similar interests. In simple words, users who liked this game also like the game too.

- Content-based Recommendations:

(See Figure 2.28)

The company must have detailed metadata (details data explanation about the data) about each of buyers. Retailers can recommend buyers with similar tags. For example, Fathimah watch the movies "Game of Thrones 1" on Netflix. This movie may have metadata tags of "War", so Netflix recommends to Fathimah another movis with metadata tags "Ancient".

6. Market basket analysis (transaction data)

Market basket analysis may be regarded as a traditional analysis in retail. This is a modeling method based on the theory that you buy certain items, you are more (or less) to buy another item. This process mainly depends on the amount of data collected via customers transactions. It works by looking for combinations of items in transactions.

- Assume there are 1000 customers
- 100 of them bought only CD games, 850 bought Playstation and 50 bought both of them.
- Bought CD games => bought playstation
- Support = P(CD games & Playstation) = 50/1000 = 0.05
- Confidence = support/P(Playstation) = 0.05/(850/1000) = 0.05/0.85=0.06
- Lift = confidence/P(CD games) = 0.75/0.10 = 7.5

Note:

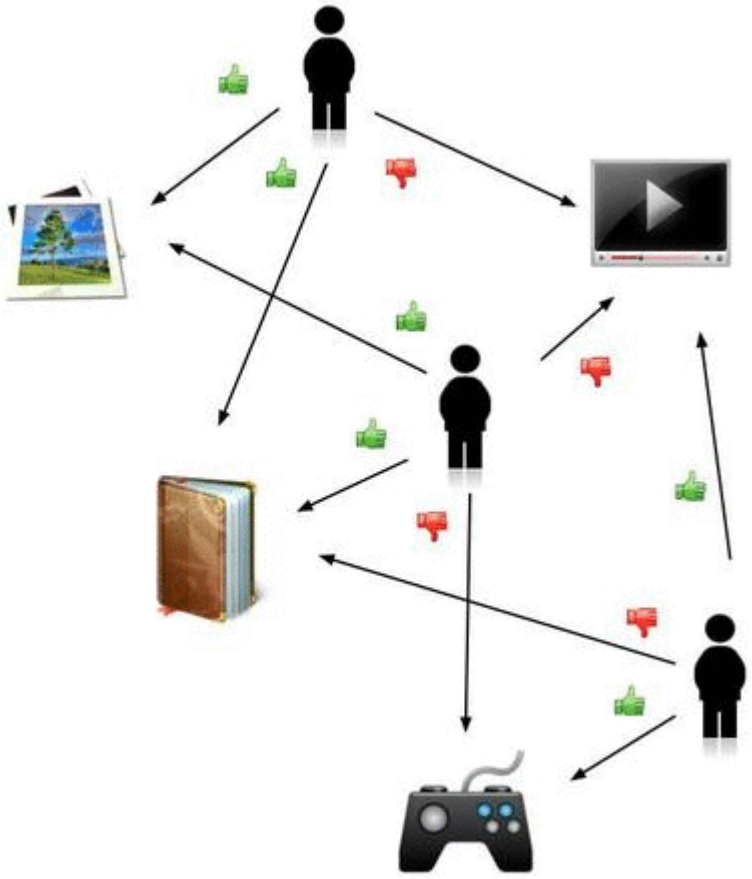


Figure 2.27:

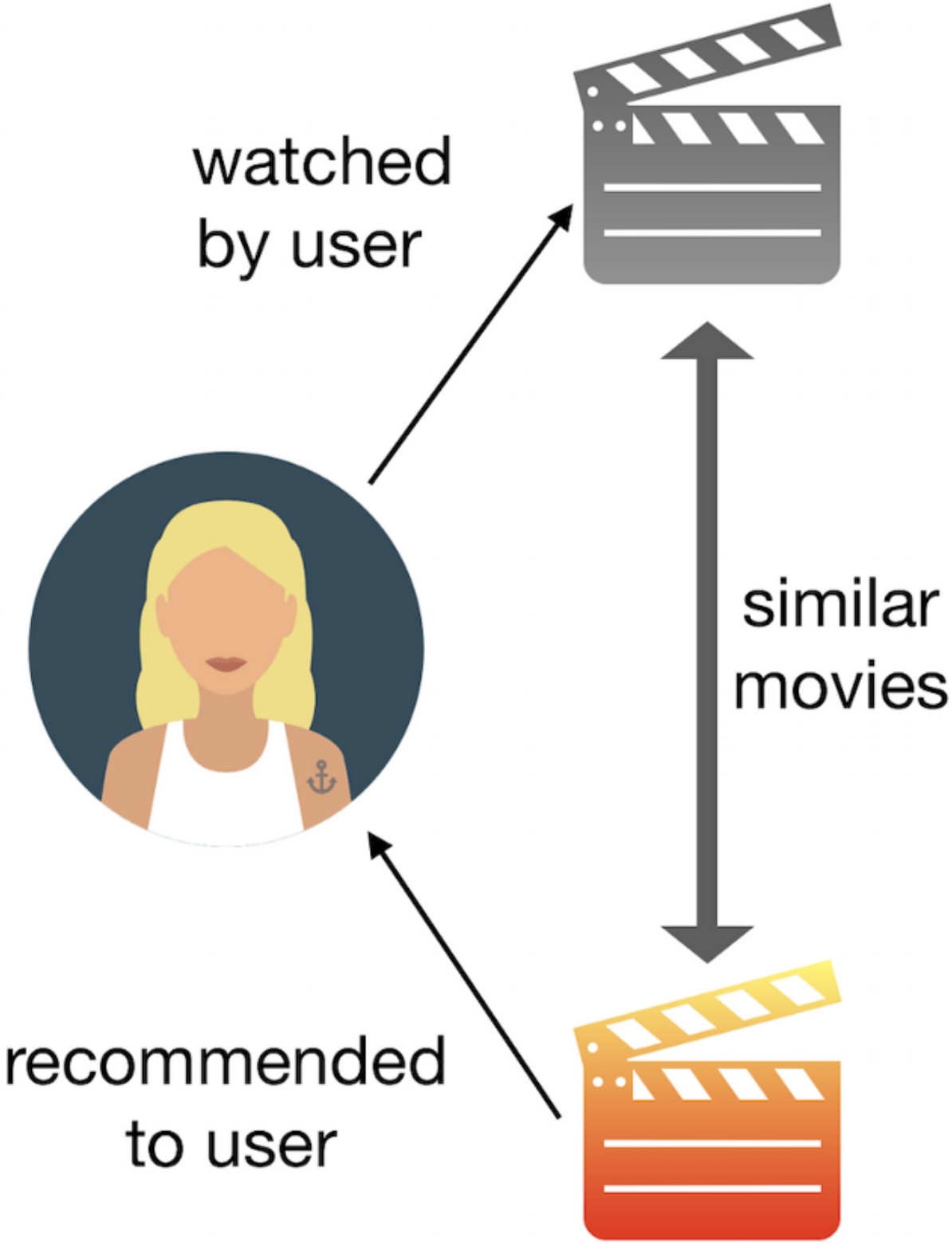


Figure 2.28:

Figure 2.29:

Support:= the number of transactions that include items in the {A} and {B} parts of the rule as a percentage of the total number of transactions.

Confidence:= the ratio of the number of transactions that include all items in {B} as well as the number of transactions that include all items in {A} to the number of transactions that include all items in {A}

lift:= the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of B

7. Fraud detection

(See Figure 2.29)

Fraud can encompass improper payments money laundering, terrorist financing, cybersecurity, and public security. The detection of fraud is an important and challenging activity of retailers. The only good way to protect retailers is to be one step ahead of fraudsters.

To identify of fraud-while improving customers experiences-retailers should follow four steps:

1. Extract and transform all available data types from across departments or channels and incorporate them into the analytical process.
2. Continually monitor transactions, social networks, high-risk anomalies, etc., and apply behavioral analytics to enable real-time decision making.
3. Install analytics culture through data visualization at all levels, including investigative workflow optimization.
4. Employ layered security techniques (another company who experts in this area).
5. Warranty analytics

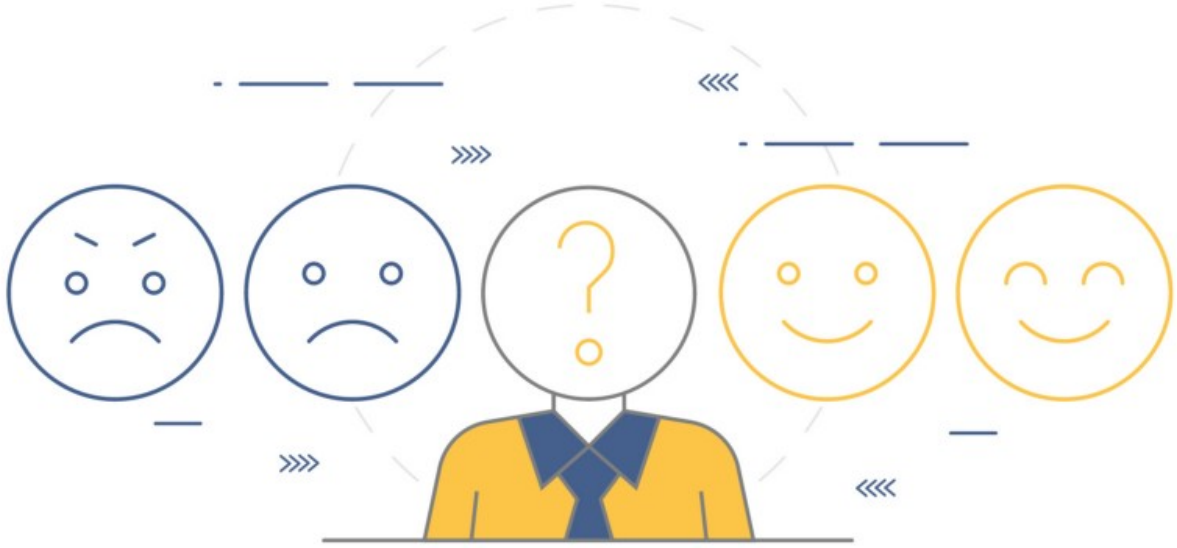


Figure 2.30:

With warranty issues costing retailers billions and indirect cost factors. To reduce warranty costs, build loyalty customer service, smart manufacturers, and focus on post-sales service. The methods of detecting are complicated. They concentrate on the detecting anomalies in the warranty claims.

The warranty analytics have some of the challenges:

- Increasing warranty costs
- New issues grow in the field for months before they are detected
- Hard to integrate and analyze
- Business analysts and executives do not have ready access to warranty data and reports.
- Difficulty integrating quality and warranty data.
- Ineffective, manually intensive processes.

9. Customer sentiment analysis

(See Figure 2.30)

Social networks (social media) and online services feedbacks can perform the sentiment analysis. Because social media data are available, also easier to make analytics on social platforms. This analysis uses language processing to know a positive or negative attitude of customer. The results are as an insight for services improvement.

The examples are:

- I do not dislike this games. (Negation handling)
- Disliking games are not really my thing. (Negation, inverted word order)
- Sometimes I really hate the Gun. (Adverbial modifies the sentiment)
- I'd really truly love going out in this weather! (Possibly sarcastic)

10. Price optimization

Pricing optimization is setting a price to maximize profits. It is driven by supply and demand data, behavioral data, competition data, etc. The right price for customers and retailers is significant advantage brought by

the optimization method. The price depends not only on the costs to produce but on the competition. The big data analysis bring this issue to the next level. The data come from many sources and define buying attitude, seasoning, competitor pricing, etc. The algorithm defines the response to the changes in prices. Using real-time optimization the retailers have the advantage to attract more customers and to build personal pricing schemes.

Below are overview four 6 pricing strategies with big data:

1. Premium pricing: Retailers set costs higher than their competitors with premium services or items.
2. Pricing for Market Penetration: Retailers offer lower prices on goods and services for the first time opened.
3. Bundle Pricing: Buy 1 get 1 pricing.
4. Psychology Pricing: For example, setting the price of a watch at $499 is proven to attract more consumers than setting it at $500.
5. Price Skimming: The retaliers offer lowers prices as competitor goods appear on the market not only for the first time opened.
6. Economy Pricing: An example of economy pricing is generic food sold at grocery stores. The generic They need little marketing and promotion expenses.is incredibly effective for large retailers companies.

11. Location of new stores

Data analytics is efficient with this issue. The method is simple and useful. The data are customers data, ZIP code, and place for understanding the market potential. The Analyst finds the solution by connection all the data source.

12. Lifetime value prediction

The customer lifetime value (LTV) is one of the most important parameters in the retail industry. CLV is a total value of customers profit to the retailer over the entire customer business relationship. Retail can measure how long it takes to retargeting the investment required to earn the customers.

The forecast is made on the past data to the most recent transactions data. The application of the statistical methodology and machine learning helps the retailer to identify customers buying pattern. It is an important business process for retailers. More customers understanding, more high LTV value.

Sources:

- `Top 10 Data Science Use Cases in Retail [11]`
- `The Impact of Big Data on The Retail Sector: Examples And Use-Cases [12]`
- `Five Big Data Use Cases for Retail [13]`
- `Collaborative filtering [14]`
- `Recommender Systems through Collaborative Filtering [15]`
- `Reduce warranty costs and improve product quality and brand reputation [16]`

2.3.4 Financial Services

Big data analytics can be the main driver of innovation in the financial industry. Using big data analytics in the financial industry is more than a trend, it has become the core analytics. The Banks and Insurances company have to realize that big data can help them more efficient, smarter decisions, and improve performance. This can be applied to the following activities:

- Banking and Insurance Industry:
  - Discovering the spending patterns of the customers
  - Identifying the main channels of transactions (ATM withdrawal, credit/debit card payments
  - Splitting the customers into segments according to their profiles
  - Product cross-selling based on the customers' segmentation
  - Fraud management & prevention

  - Risk assessment, compliance & reporting
  - Customer feedback analysis and application

- Only Insurance Industry:

  - Better product design and marketing
  - More accurate risk assessment, underwriting, and pricing processes
  - Stronger commitment to helping customers

- Only Banking Industry:

  - Better claims management
  - Better customer targeting and ensuring growth
  - Enhancing risk assessment
  - Improving productivity and decision-making
  - More business opportunities
  - Digital banks – internet-based banks

Here is a list of big data use cases in banking area:

1. Fraud detection

Machine learning is effective detection and prevention of fraud include credit cards, accounting, insurance, and more. The key steps to fraud detection include: Obtaining data sampling for model estimation and preliminary model testing.

The fraud detection needs deep theoretical knowledge into practical applications. Several algorithms are needed, such as association, clustering, forecasting, and classification. The simple example of efficient fraud detection is outlier data. When some unusually high transactions found or it can investigate the unusually high purchase of popular items and multiple accounts.

Palmer (2014) suggests that the following stages of enterprise counter-fraud measures:

- **Detect:** By applying advanced analytics to all key fraud data to aid in predicting if an action is potentially fraudulent.
- **Respond:** Applying fraud insights to take action in real time.
- **Investigate:** By performing and managing inquiries into a suspicious activity that are supported by data analysis and collaborative sophisticated case management.
- **Discover:** Using new big data analytics capabilities that help in the identification of suspicious activity by analyzing historical data to search for patterns of fraud and financial crimes.

2. Marketing

Xerago (2015) defines marketing analytics as the practice of measuring, managing and analyzing market performance to maximize the effectiveness of and return on investment (ROI) from the marketing activities. Marketing analytics will help banks through data to increase profitability.

Pramanick (2013) states that banks are always at risk of losing customers and need strategies that are dependent on identifying the right action to the right customer. Thus banks need customer analytics to segment the customers. This marketing will assist in determining products and services, pricing, and strategy of the banks.

Morabito (2015) adds that big data-enabled marketing automation will assist banks in servicing individual customer needs while keeping the marketing costs low, enabling a personalized experience at a good ROI.

3. Credit Risk Management

Sas (sas.com: one of the leading analytics software company) defines credit risks the probability of loss due to a borrower's failure to make payments on any type of debt. Better credit risk presents an opportunity and improve performance and also a competitive advantage.

Due to the SAS Institute (sas.com), credit risk have also challenges, such as:

- **Inefficient data management**: An inability to access the right data when it's needed causes problematic delays.
- **No groupware risk modeling framework**: Without it, banks can't generate complex, meaningful risk measures and get a big picture of groupware risk.
- **Constant rework**: Analysts can't change model parameters easily, which results in too much duplication of effort and negatively affects a bank's efficiency ratio.
- **Insufficient risk tools**: Without a robust risk solution, banks can't identify portfolio concentrations or re-grade portfolios often enough to effectively manage risk.
- **Cumbersome reporting**: Manual, spreadsheet-based reporting processes overburden analysts and IT.

The solution should include:

- Better model management that spans the entire modeling life cycle.
- Real-time scoring and limits monitoring.
- Robust stress-testing capabilities.
- Data visualization capabilities and business intelligence tools that get important information into the hands of those who need it, when they need it.

Sources:

- `Big Data analytics in the banking sector [17]`
- `How to derive value from big data [18]`
- `The Impact of Big Data Analytics on the Banking Industry [19]`
- `Data Analytics in the Financial Services Industry [20]`

2.3.5 Healthcare

(See Figure 2.31 and Figure 2.32)

The industry has taken advantage of big data and analytics to make strategic business decisions. Healthcare big data refers to the quantities of data that is now available to healthcare providers. The big data in the healthcare industry is changing the way patients and doctors handle care. The more big data involved, the more efficient healthcare can be.

Benefits of Healthcare Big Data:

- Create holistic, 360-degree views of consumers, patients, and physicians.
- Improve care personalization and efficiency with comprehensive patient profiles.
- Inform physician relationship management efforts by tracking physician preferences, referrals, and clinical appointment data.
- Boost healthcare marketing efforts with information about consumer, patient, and physician needs and preferences.
- Analyze trends within a single hospital or greater a healthcare network to benefit research and care procedures for enhanced population health outcomes.
- Identify patterns in health outcomes, patient satisfaction, and hospital organization.
- Predict health outcomes and create preventive care strategies with data analysis.
- Optimize growth by improving care efficiency, effectiveness, and personalization.

Below are common questions about Healthcare Big Data:

- What is Driving the Adoption of Big Data in Healthcare?
  - Volume of Healthcare Data
  - Government Regulations

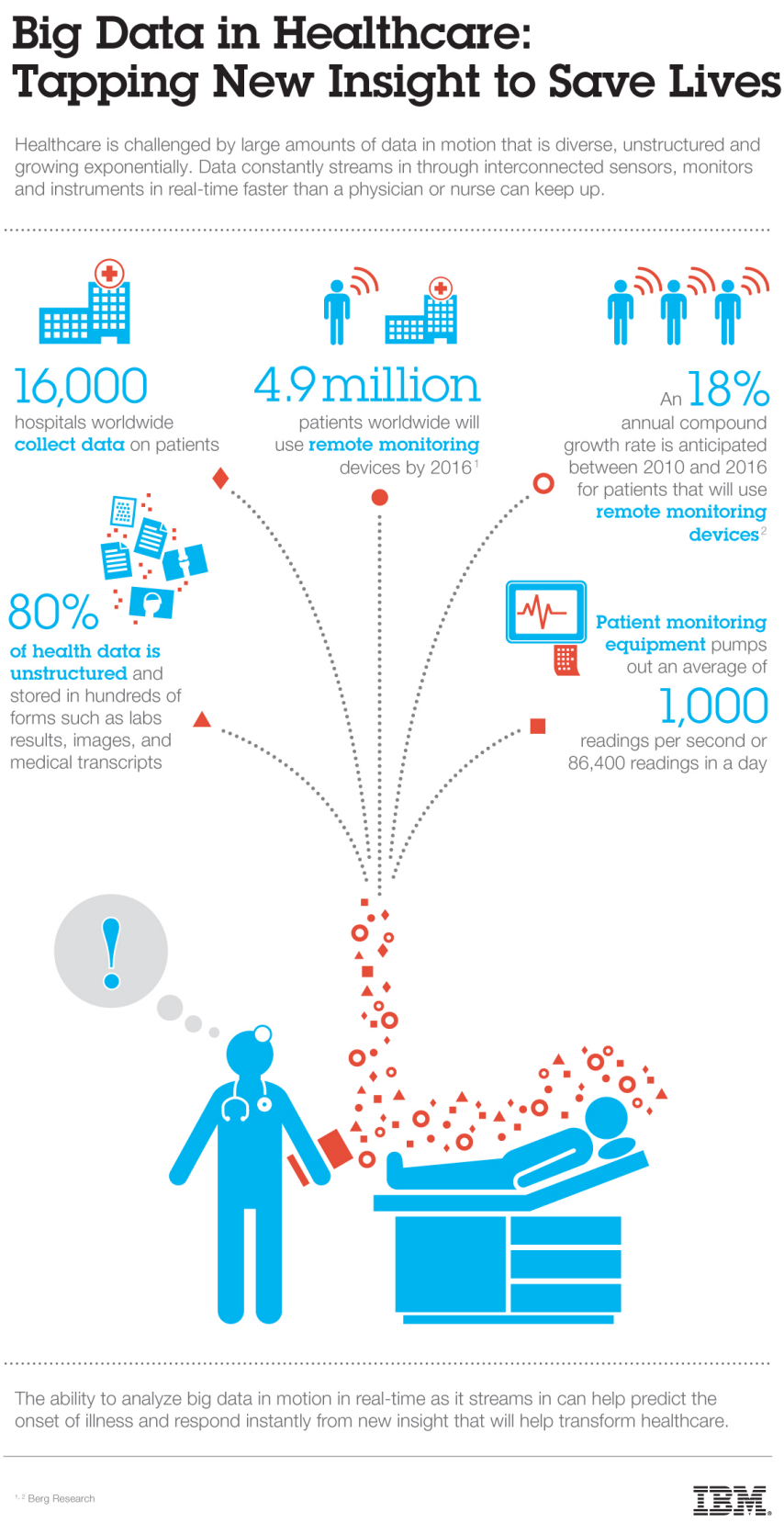

Big Data in Healthcare:
Tapping New Insight to Save Lives
Healthcare is challenged by large amounts of data in motion that is diverse, unstructured and growing exponentially. Data constantly streams in through interconnected sensors, monitors and instruments in real-time faster than a physician or nurse can keep up.
16,000
hospitals worldwide collect data on patients
4.9 million
patients worldwide will use remote monitoring devices by 2016[1]
An 18%
annual compound growth rate is anticipated between 2010 and 2016 for patients that will use remote monitoring devices[2]
80%
of health data is unstructured and stored in hundreds of forms such as labs results, images, and medical transcripts
Patient monitoring equipment pumps out an average of
1,000
readings per second or 86,400 readings in a day
The ability to analyze big data in motion in real-time as it streams in can help predict the onset of illness and respond instantly from new insight that will help transform healthcare.
[1,2] Berg Research
IBM

Figure 2.31:

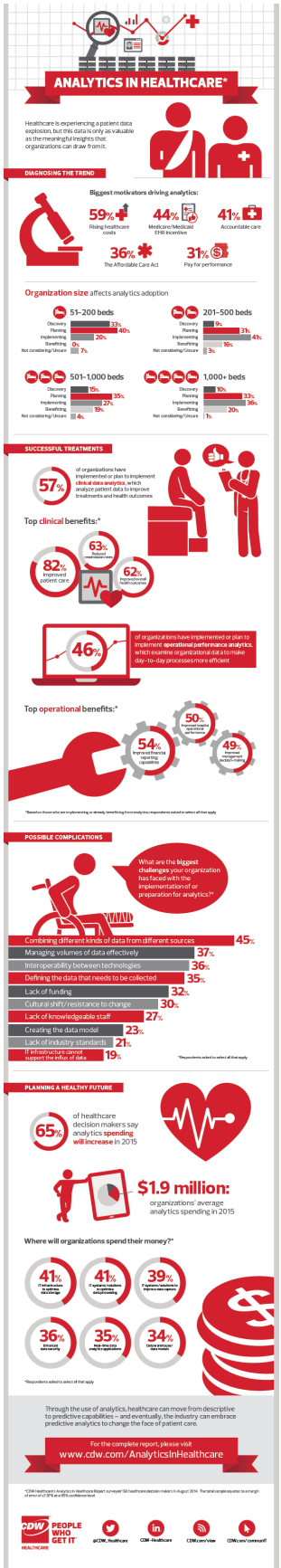

ANALYTICS IN HEALTHCARE*
Healthcare is experiencing a patient data explosion, but this data is only as valuable as the meaningful insights that organizations can draw from it.
DIAGNOSING THE TREND
Biggest motivators driving analytics:
59% Rising healthcare costs
44% Medicare/Medicaid EHR incentive
41% Accountable care
36% The Affordable Care Act
31% Pay for performance
Organization size affects analytics adoption
51–200 beds
201–500 beds
501–1,000 beds
1,000+ beds
Discovery
Planning
Implementing
Benefiting
Not considering/Unsure
SUCCESSFUL TREATMENTS
57% of organizations have implemented or plan to implement clinical data analytics, which analyze patient data to improve treatments and health outcomes
Top clinical benefits:*
82% Improved patient care
63% Reduced readmission rates
62% Improved overall health outcomes
46% of organizations have implemented or plan to implement operational performance analytics, which examine organizational data to make day-to-day processes more efficient
Top operational benefits:*
54% Improved financial reporting capabilities
50% Improved clinical and financial performance
49% Improved management decision-making
*Based on those who are implementing or already benefiting from analytics; respondents asked to select all that apply
POSSIBLE COMPLICATIONS
What are the biggest challenges your organization has faced with the implementation of or preparation for analytics?*
Combining different kinds of data from different sources 45%
Managing volumes of data effectively 37%
Interoperability between technologies 36%
Defining the data that needs to be collected 35%
Lack of funding 32%
Cultural shift/resistance to change 30%
Lack of knowledgeable staff 27%
Creating the data model 23%
Lack of industry standards 21%
IT infrastructure cannot support the influx of data 19%
*Respondents asked to select all that apply
PLANNING A HEALTHY FUTURE
65% of healthcare decision makers say analytics spending will increase in 2015
$1.9 million: organizations' average analytics spending in 2015
Where will organizations spend their money?*
41%
41%
39%
36%
35%
34%
*Respondents asked to select all that apply
Through the use of analytics, healthcare can move from descriptive to predictive capabilities – and eventually, the industry can embrace predictive analytics to change the face of patient care.
For the complete report, please visit
www.cdw.com/AnalyticsInHealthcare
*CDW Healthcare's Analytics in Healthcare Report surveyed 150 healthcare decision makers in August 2014. The total sample equates to a margin of error of ±7.97% at a 95% confidence level
CDW PEOPLE WHO GET IT
HEALTHCARE
@CDW_Healthcare
CDW - Healthcare
CDW.com/View
CDW.com/connect

Figure 2.32:

  - Desire for Personalized Care
- How Can Marketing Take Advantage of Healthcare Big Data?
  - Propensity models (score potential targets)
  - Communication personalization (continue an ongoing relationship with the healthcare organization)
  - Integrated communication (holistic customer experiences throughout the care continuum)
- What Challenges Arise with Healthcare Big Data?
  - Sorting and prioritizing of information
  - Ensuring that the right access to big data insights and analysis (data security issues)
- What is the Future of Healthcare Big Data?
  - Big data are more crucial for company success
  - Big data healthcare will also smarter and more integrated
  - Big data will grow with Internet of Things (IoT)
  - Iot will become standard methods

Sources:

- `The Powerful Role of Big Data In The Healthcare Industry [21]`
- `12 Examples of Big Data Analytics In Healthcare That Can Save People [22]`

2.3.6 Crime & Terorism

The use of fingerprints, DNA, ballistic analysis, CCTV and other types of technology have also played to improve big data analytics in crime. A top use case for crime is a predictive analytics. This helps in this aspect:

- Police can make compelling cases to get emergency resources to fight recent crime waves
- Police can identify the likelihood that they are dealing with serial offenders
- Police can look for precipitating factors that cause crime epidemics and pass that information along to policymakers to take preventive measures

The success story of crime prediction come from RUSI (Royal United Services Institute for Defence and Security Studies). They published a paper "Big Data and Policing An Assessment of Law Enforcement Requirements, Expectations, and Priorities". One to the successful project is **Predictive Crime Mapping**

Sources:

- `Police Are Using Big Data To Predict Future Crime Rates [23]`
- `Big Data and Policing An Assessment of Law Enforcement Requirements, Expectations and Priorities [24]`

Chapter 3

Big Data Analytics

3.1 Data Analytics Lifecycle

One of the most methods for data analytics is **CRISP-DM methodology.** The method was found by five companies: SPSS, Teradata, Daimler AG, NCR Corporation, and OHRA (an insurance company). The CRISP-DM methodology stands for Cross Industry Standard Process for Data Mining. It is a robust and well-proven methodology. This model is an idealized iterative method. Take a look at the following illustration. See (Figure 3.1)

Stage One- Determine Business Objectives

This first phase focuses on understanding the project goals and requirements from a business perspective. The goal of this stage is to uncover important factors for improving business. Neglecting this step can mean that a great deal of effort is put into producing the right answers to the wrong questions.

Stage Two - Data Understanding

Accessing and exploring data using statistical and mathematical methods. It includes describe data, explores data, verifies data quality, and data quality report. The name of this steps is exploratory data analysis (EDA).

1. Describe data: Sources of a dataset, quantity, structured of the data (structured or unstructured data), etc.
2. Explore data: It may include,

- Distribution of attributes (for example, the target attribute of a prediction task)
- Relationships between attributes of the dataset
- Results of aggregations (SUM, AVerage)
- Simple statistical analyses (MEdian, MAX, MIN)

3. Verify data quality: Identify the quality of the data, it can be,

- Data completeness
- Data correctness, or does it contain errors and, if there are errors, how common are they?
- The missing values.

4. Data quality report: Complete Report about data quality. There are many methods and software doing this steps. For example with R Programming. See the next chapter: `4.2 EDA: Exploratory Data Analysis`

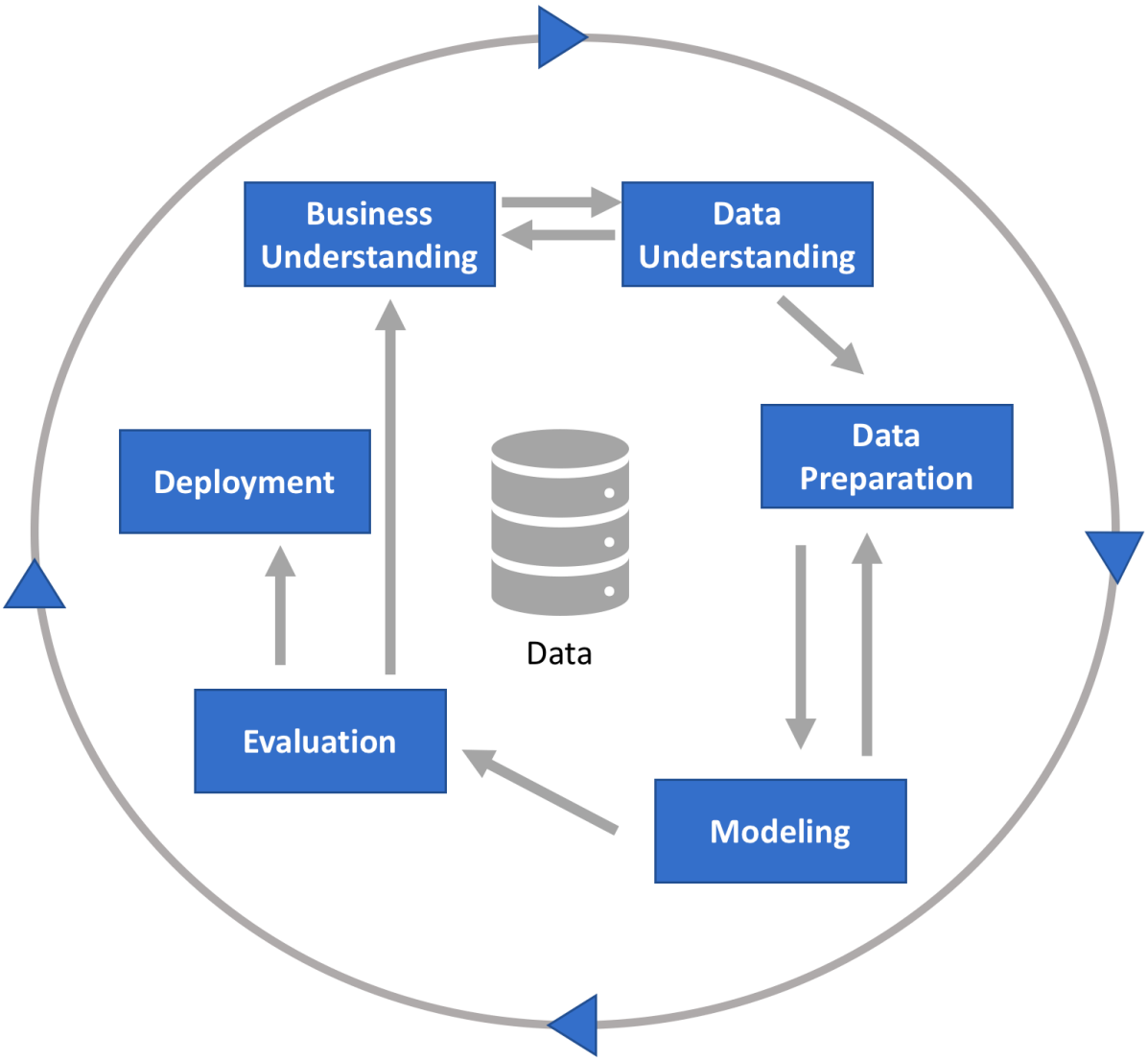

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Figure 3.1:

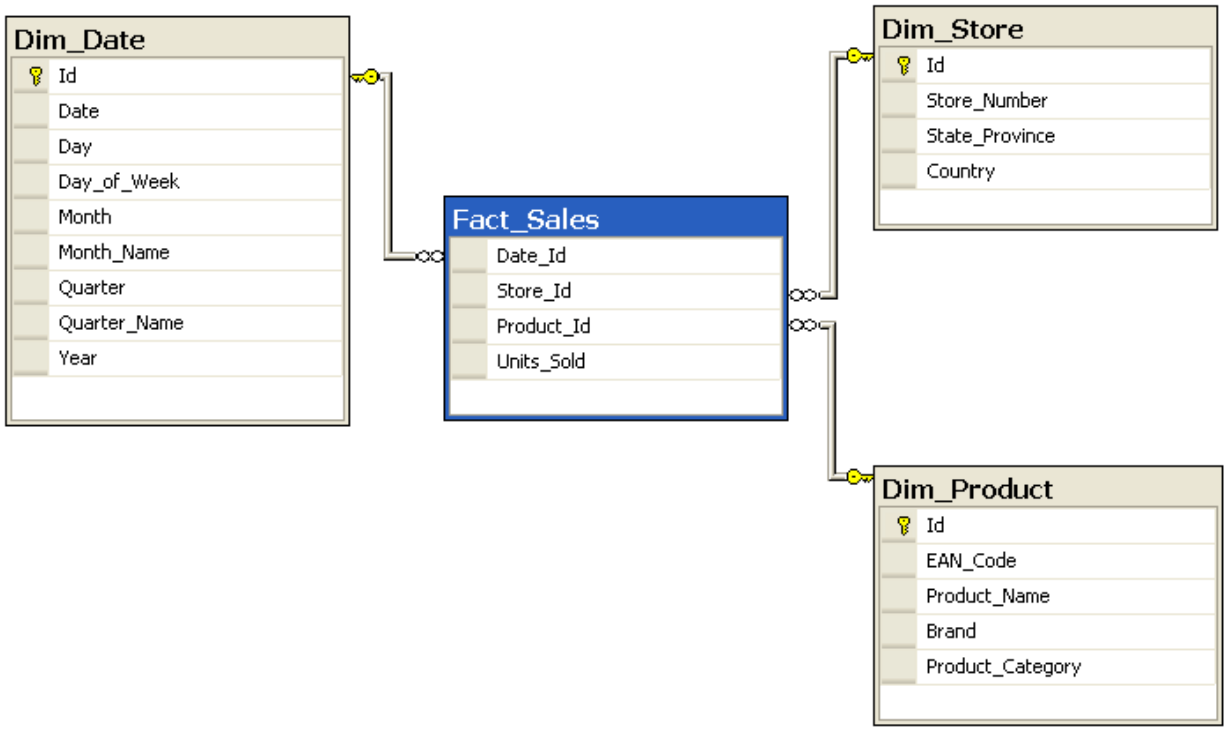


Figure 3.2:

Stage Three - Data Preparation

In this phase, various data sources are selected to construct the final dataset. The Tasks include data cleaning, attribut seelction, and data transformation. Below are steps by steps in this stage. See the next chapter about data cleaning in R Programming `4.1 Data Cleaning`

1. Select your data
2. Clean your data
3. Construct required data
4. Integrate data

Stage Four - Data Modeling

What is Data Modelling?: Data modelling is the process to create final dataset for the stored in a database or another type storage systems (non relational noSQL).

Why use Data Model? The aim of this phase is to describe:

- The data contained in the database (e.g., entities: students, lecturers, courses, subjects, note)
- The relationships between data items (e.g., students are supervised by lecturers; lecturers teach courses)
- The attribute on data (e.g., student number has exactly eight digits; a subject has four or six units of credit only)

There are many model in this terms. But the two famous are `Star schema` and `Snowflakes schema`.

`Star schema`

The star schema consists of two types of table. Fact table and Dimension table. (See Figure 3.2)

`Snowflakes schema`

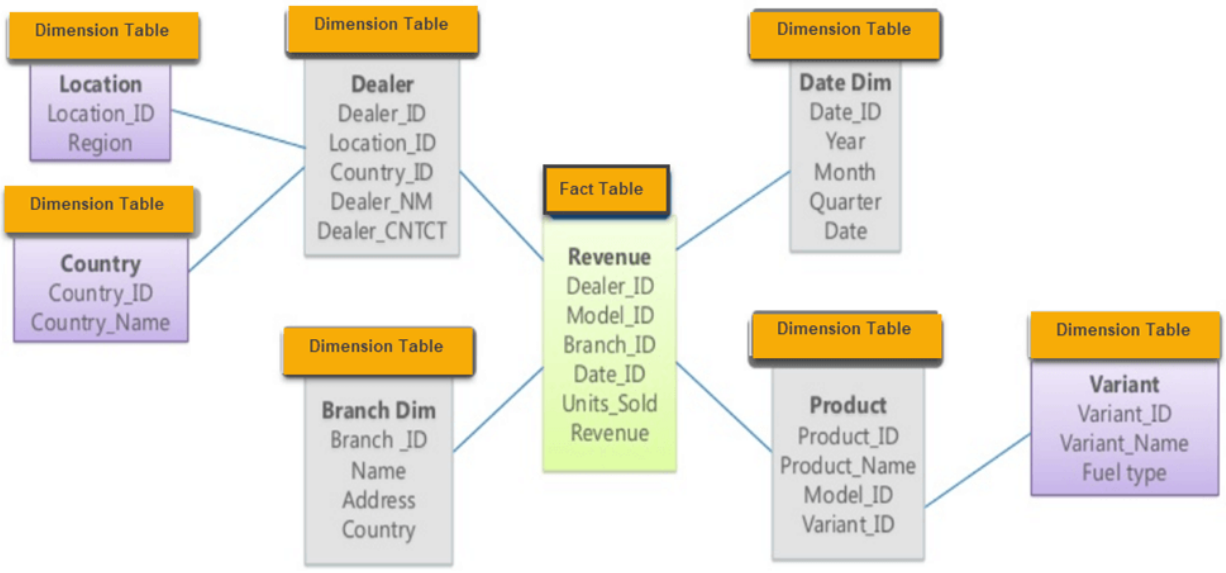


Figure 3.3:

Star and snowflake schemas are similar at heart. A Snowflake Schema is a modification of a Star schema. The main different from Star schema are hierarchies divided into separate tables like Figure 3.3.

Stage Five - Evaluation

The accuracy and generality of the model is the key to this step. A key objective is to determine if there is some important business point or some other data sources that have not been sufficiently considered.

Stage Six - Deployment

The substeps of this phase are:

1. Plan deployment
2. Plan monitoring and maintenance
3. Produce final report
4. Review project

Note: The CRISP-DM is iterative process. It meants a systematic, repetitive, and recursive process.

For more info and sources:

- `What is the CRISP-DM methodology? [25]`
- `Big Data Analytics Tutorial [26]`
- `Database Design [27]`
- `Multidimensional models [28]`
- `Microsoft Intune Data Warehouse data model [29]`
- `Understanding Star Schemas [30]`

3.2 Analytics Paradigm

Analytics helps the company answer the business questions. What happened in the past, what happens in the future, and what decisions a company can make to improving future performance. Nowadays, data analytics is very important and more and more universities and further education programs are creating analytics learning environments. `Tom Davenport` (http://www.tomdavenport.com/) from CIO (Chief Information

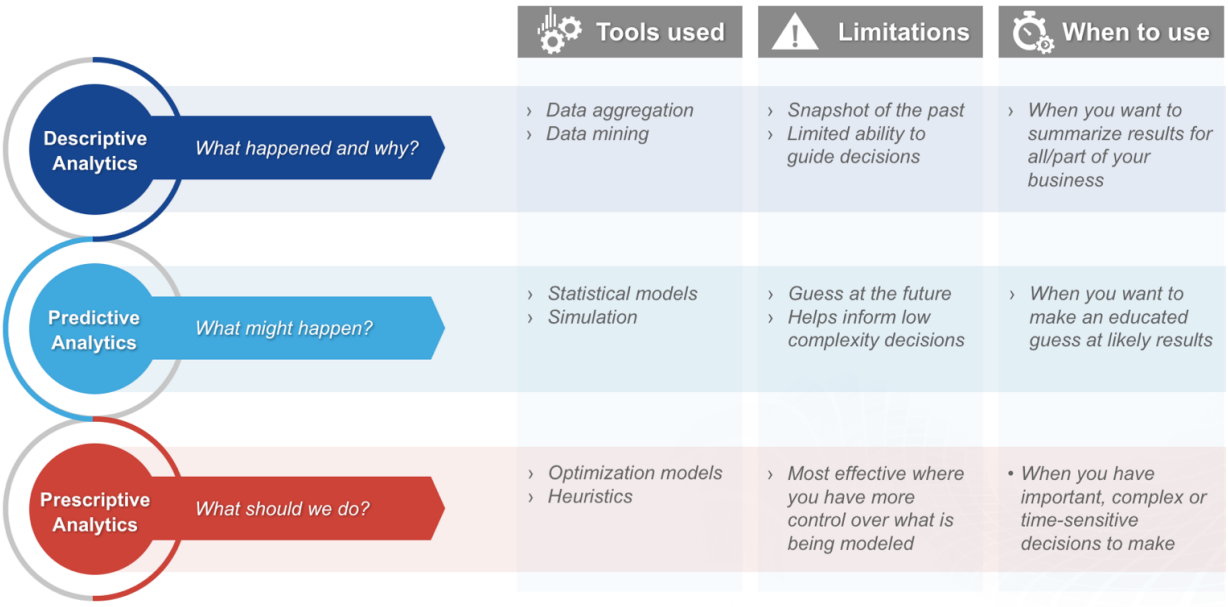


Figure 3.4:

Officer) Magazine (https://www.cio.com/) divided analytics area into three main types of analytics (See Figure 3.4). The three main components of analytics are as follows:

Descriptive Analytics:

Descriptive Analytics helps the company to get insight from the past and current state of the business.

Most every business function in the company (Production, Sales, Finance, etc.) used Descriptive Analytics to create custom reports.

Overview:

- What happened and why?
- Tools and method: Data aggregation and data mining
- Insight of the past with limited ability to help decisions process
- To summarize results of your business

Predictive Analytics:

Predictive Analytics predicts the business situation in the future. It need applies statistical techniques (often machine learning or Artifical Intelligence) to get what the future may be happen.

Overview:

- What might happen?
- Tools and methods: Machine learning, statistical models, and simulation (Risk Management Simulation)
- Limitation to be aware of: Guess at the future, helps inform low complexity decisions
- When teh company want to make an educated predictions at likely results

Prescriptive Analytics:

It optimized the set of decisions. Company can quickly evaluate trillions or more possible combinations of choices (for example: what products in which manufacturing, what product lines and in what quantities), minimum production of a given product, manufacturing time and cost, machine, raw material inventory,

finished goods inventory capacity, and maximize or minimize your objectives (for example: total product costs).

Overview:

- What should the company do?
- Methods: Optimization (Linear and Nonlinear)
- Limitation to be aware of: Most effective where the company have some control over what is being modeled
- When the company have important, interdependent, complex or time-sensitive decisions to make

Source:

- `Descriptive, Predictive and Prescriptive Analytics [31]`

3.3 Deutsche Bahn Tools

3.3.1 SAP Business Object Web Intelligence

SAP BO Web Intelligence is databases system created by SAP. The reason for choosing this databases system is robust software (from a few users to tens of thousands of users) and excellent service from SAP company. This SAP system has some other advantages, such as:

- On-premise deployment (offline)
- Real-time business intelligence (offline and offline)
- Increased user autonomy
- Make information consumption simple, personalized, and dynamic

More about SAP Business Object Web Intelligence: https://www.sap.com/products/bi-platform.html

3.3.2 Open Refine

Openrefine is a data manipulation tool. Openrefine can cleans, reshapes messy and unstructured data. OpenRefine provides the flexibility to choose dataset from a variety of data sources (database, website, Excel, Microsoft Access, Web Based Software, ect). Users can use this tool to get a big view of their dataset in terms of statistically insight.

What – A messy, unstructured, inconsistent dataset can be explored. OpenRefine gives several functions such as filter the data, edit the inconsistencies, and view the data. It's a tool to clean the messy data.

Why - Openrefine cleans the data in a more systematic controlled manner than others tools. All of the change in the dataset is recorded.

When -The Data analyst process: Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time data cleaning which is an inefficient data strategy.

Why OpenRefine is a better tool for data cleaning and modeling:

Google refine	Spreadsheets	Databases
Batch editing of rows and columns possible	Editing of one cell at a time	Schema and programming language required for editing
Used for exploring and transforming data	Used for entering data and performing calculations, functions	Data is out of sight unless script is run to view it.
No Schema Required	No Schema Required	
Data is always visible at each step of editing	Data is always visible	
More interactive and visual	Visual is not impressive.	

Sources and tutorial:

- `Using OpenRefine [32]`
- `Open Refine Official Site: http://openrefine.org/`
- `Tutorial: OpenRefine [33]`

3.3.3 R Programming

Key points R Programming:

- An Open Source programming
- A Programming for statistical computing and graphics.
- It maintained by the R core volunteer developers
- Founded by `Ross Ihaka` and `Robert Gentlemen`.

More info about R: https://www.r-project.org/

3.3.4 RStudio

Key points RStudio:

- An Open Source IDE (Integrated Development Environment) for R programming
- `JJ Allaire` found RStudio (JJ Allaire is the creator of the programming language ColdFusion).
- It runs on the desktop and browser (Windows, macOS, and Linux, Debian, Ubuntu, Red Hat Linux, CentOS, openSUSE and SLES).

More info about RStudio: https://www.rstudio.com/

3.3.5 Microsoft Power BI

1. What Is Power BI?
2. Why Power BI?
3. Who Use Power BI?
4. Components Of Power BI
5. What Is Power BI?

Based on powerbi.microsoft.com. Power BI is a Data Visualization and Business Intelligence tool that converts data from different data sources to interactive dashboards and reports. For making best business decisions, it is important to get relevant information from the data sources and present it as visualization. Power BI has several products, such as:

- Power BI Desktop: Create rich, interactive reports with visual analytics at your fingertips—for free in desktop application.
- Power BI Pro: Connect to hundreds of data sources, and visualize all your data with live dashboards and reports. Then share insights across your organization to fuel intelligent action.
- Power BI Premium: Power BI Premium offers advanced, self-service data preparation that allows every user—from business analyst to data scientist—to accelerate the delivery of insights and collaborate with ease.
- Power BI Mobile: Stay connected to your data wherever business takes you. Mobile business intelligence is just a touch away.
- Power BI Embedded: Embed interactive Power BI visuals in your applications, websites, or portals to bring world-class analytics directly to your customers
- Power BI Report Server: Power BI Report Server is the on-premises solution for reporting today, with the flexibility to move to the cloud tomorrow. It's included with Power BI Premium so you have the ability to move to the cloud on your terms.

2. Why Power BI?

- Power BI is built on the convention of best BI products available – SQL Server Analysis Services (SSAS) and Microsoft Excel. Power BI gives you the option to connect with these products (heavy users Microsoft products).
- Power BI is Being built/rebuilt using the latest technologies like HTML 5.0, cloud computing, column store databases & smartphone mobile apps.
- Microsoft has opened the custom visuals gallery for open source contributions which adds value to the community (from users and another company as adds in).
  - https://community.powerbi.com/t5/Data-Stories-Gallery/bd-p/DataStoriesGallery
  - https://appsource.microsoft.com/en-us/marketplace/apps?product=power-bi-visuals
- Trend towards self-service business intelligence indicates Microsoft's leading position in this space. Based on Gartner Research (the world's leading research and advisory company) Power BI get high position. (See Figure 3.5)

3. Who Use Power BI?

- Developers
- IT Professional
- Subject Experts (Financial, Insurance, IT Industry, etc)
- Buiness Analyst
- Data Analyst
- Data Scienctist
- Business Intelligence Developer

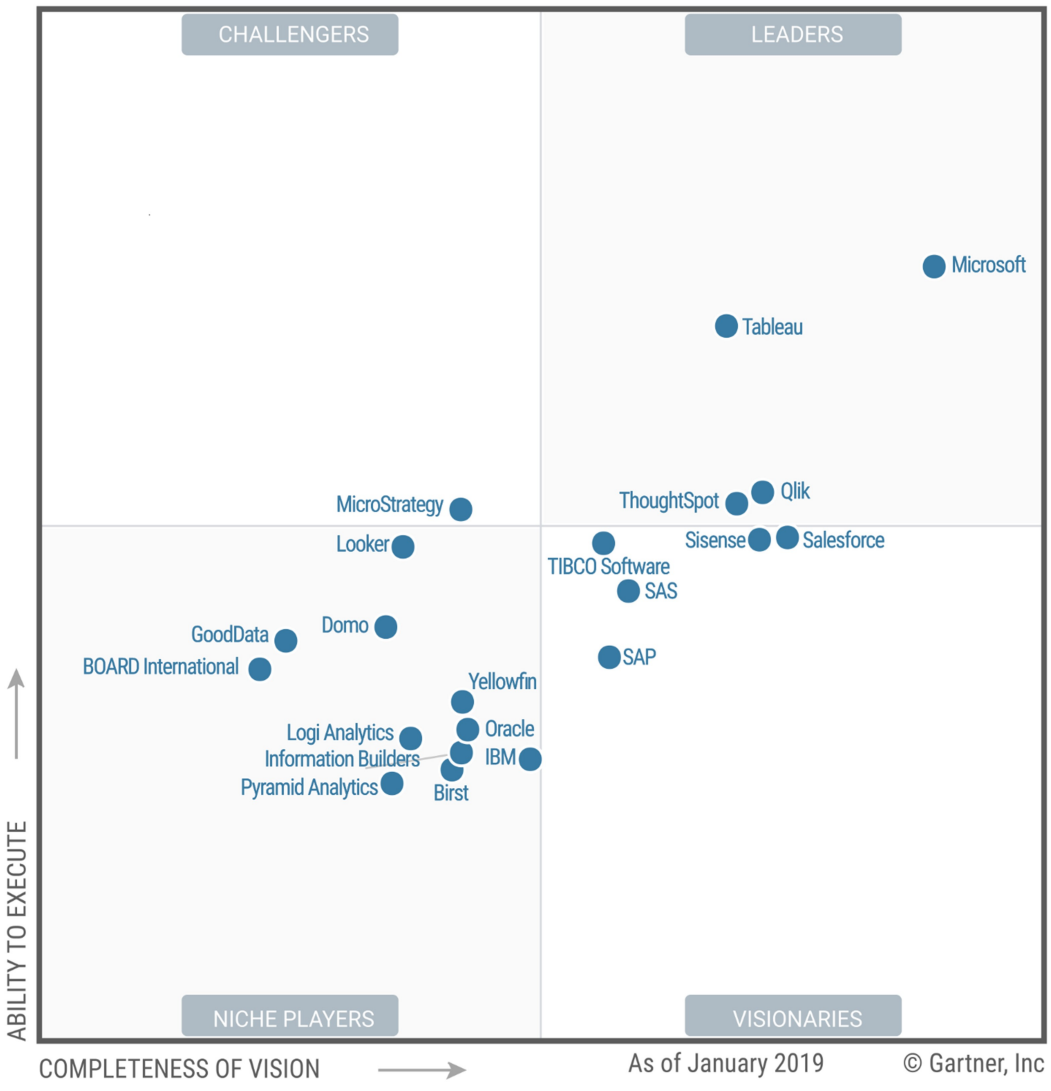
4. Components Of Power BI

- Power Query: ETL (Extract, Transform, Loading) data from different sources. Power Query Data Sources are:
  - Web page
  - Excel or CSV file
  - XML file
  - Text file
  - Folder
  - SQL Server database
  - Microsoft Azure SQL Database
  - Access database
  - Oracle database
  - IBM DB2 database
  - MySQL database
  - PostgreSQL Database
  - Sybase Database
  - Teradata Database
  - SharePoint List
  - OData feed
  - Microsoft Azure Marketplace
  - Hadoop File (HDFS)
  - Microsoft Azure HDInsight
  - Microsoft Azure Table Storage
  - Active Directory
  - Microsoft Exchange
  - Facebook

More info about Power Query: https://docs.microsoft.com/de-de/powerquery-m/power-query-m-reference

(See Figure 3.6)

- Power Pivot: Used in data modeling strategy (Snow flakes, Relational, Star Schema, etc). Power Pivot



Source: Gartner (February 2019)

Figure 3.5:

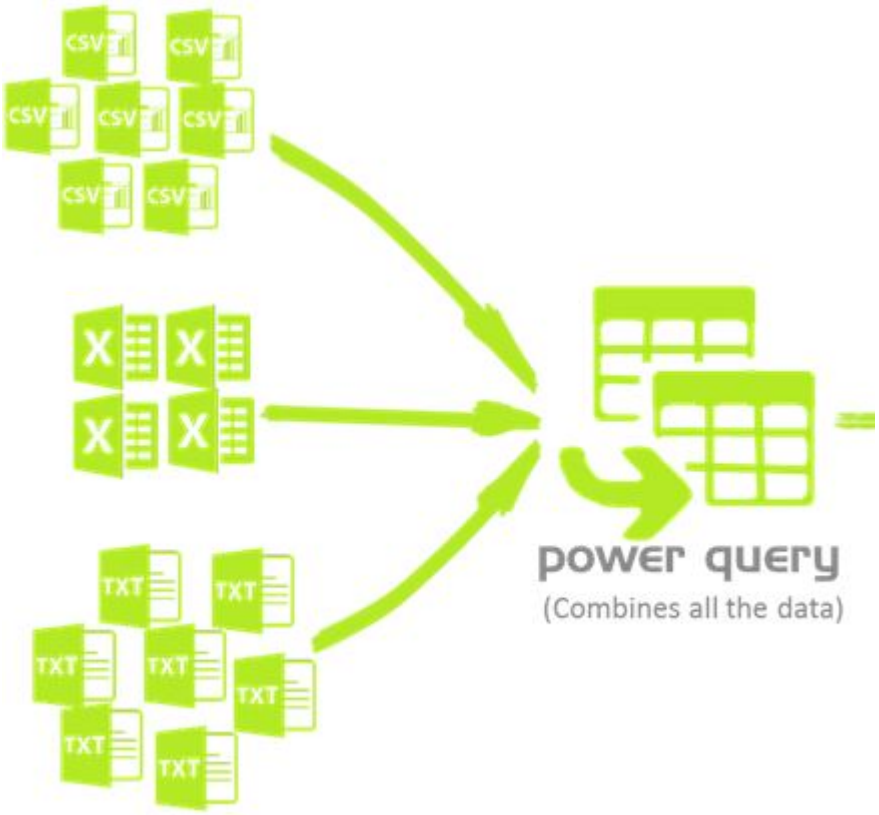

CSV
CSV
CSV
CSV
CSV
CSV
CSV
CSV
X
X
X
X
TXT
TXT
TXT
TXT
TXT
TXT
TXT
power query
(Combines all the data)

Figure 3.6:

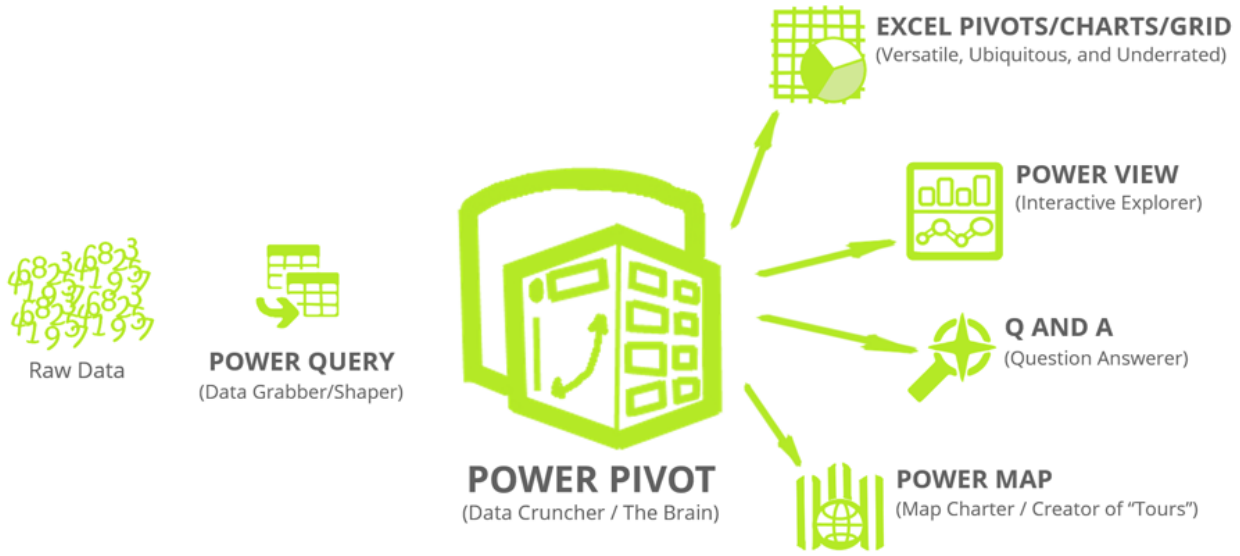


Figure 3.7:

have DAX (Data Analysis Expressions) as the function programming to help create data model from different sources. DAX is a collection of functions, operators, and constants that can be used in a formula, or expression, to calculate and return one or more values.

More info about Power Pivot and DAX:

- https://docs.microsoft.com/en-us/power-bi/desktop-quickstart-learn-dax-basics
- https://docs.microsoft.com/en-us/dax/dax-function-reference
- Power View: Analyze, visualize and display data as an interactive data visualization.
- Power Map: Interactive geographical visualization.
- Power BI Service: Share data visualization with web browser (powerbi.com)
- Power BI Q&A: Ask questions and get immediate answers with natural language query.
- Cortana for Power BI: Cortana is a voice virtual assistant created by Microsoft for Power BI.

(See Figure 3.7)

Source:

- `Edureka blog [34]`

3.3.6 Github

Github is an open source version control. Version control is a system that records changes to a file or set of files over time so that the user can recall specific versions later. See the Github site: https://github.com/

Key Points of Github as Distributed Version Control:

- Keep tracks of changes over time
- Allows the progress and projects to track
- Allows the company to revert to earlier versions
- It easier to collaborate with teams
- Track changes on various files (more than one file)

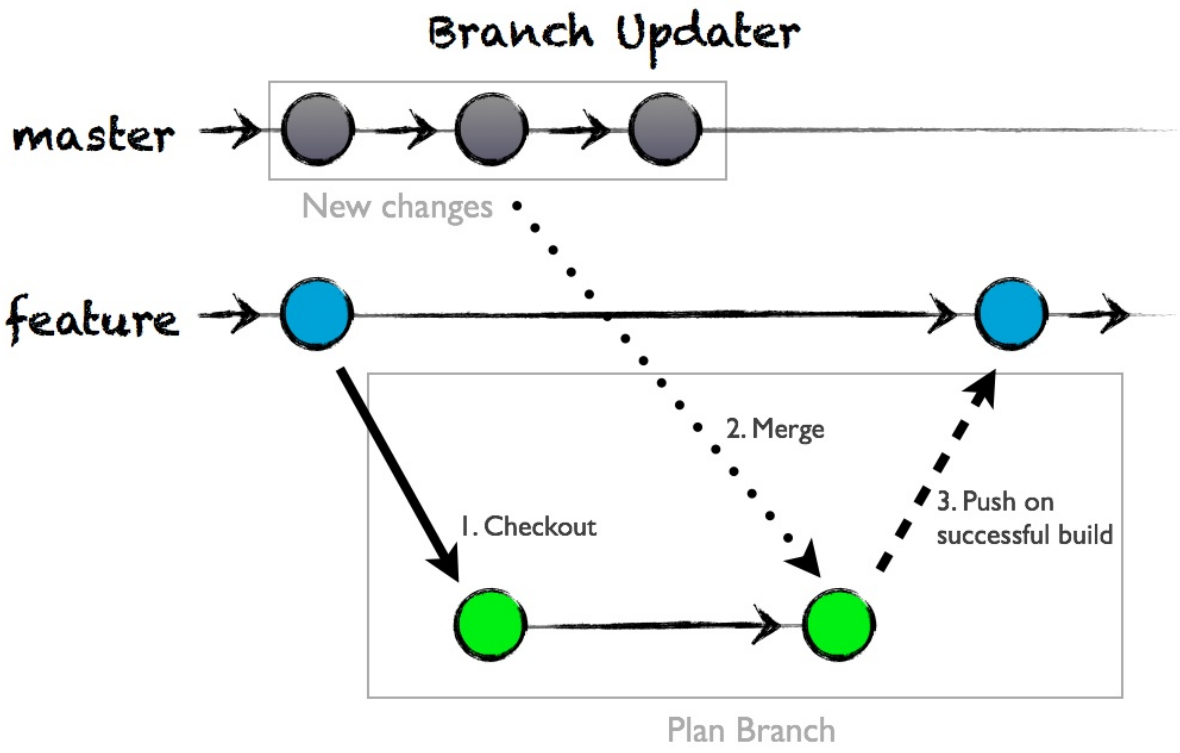


Figure 3.8:

- Track changes on a directory (Track ID)
- Allows savings non-text files (i.e. images, sheet file, etc)
- Various teams working on the same file
- Use of remote repositories to collaborate

DB use Github with RStudio. For more info about how to using Github with R: https://happygitwithr.com/

(See Figure 3.8 and Figure 3.9)

3.3.7 Confluence (Software)

Based on Wikipedia and confluence official site. Confluence is a content collaboration tool. Confluence developed and published by Australian software company Atlassian (atlassian.com). It used to help the company to collaborate and share knowledge efficiently. With Confluence, teams can create pages and blogs which can be commented on and edited by all members of the team. It is like Microsoft word with additional features.

- More information: https://www.atlassian.com/software/confluence
- Confluence have about 180 features as documentation cloud software. Here the complete list: https://bit.ly/2CADurI

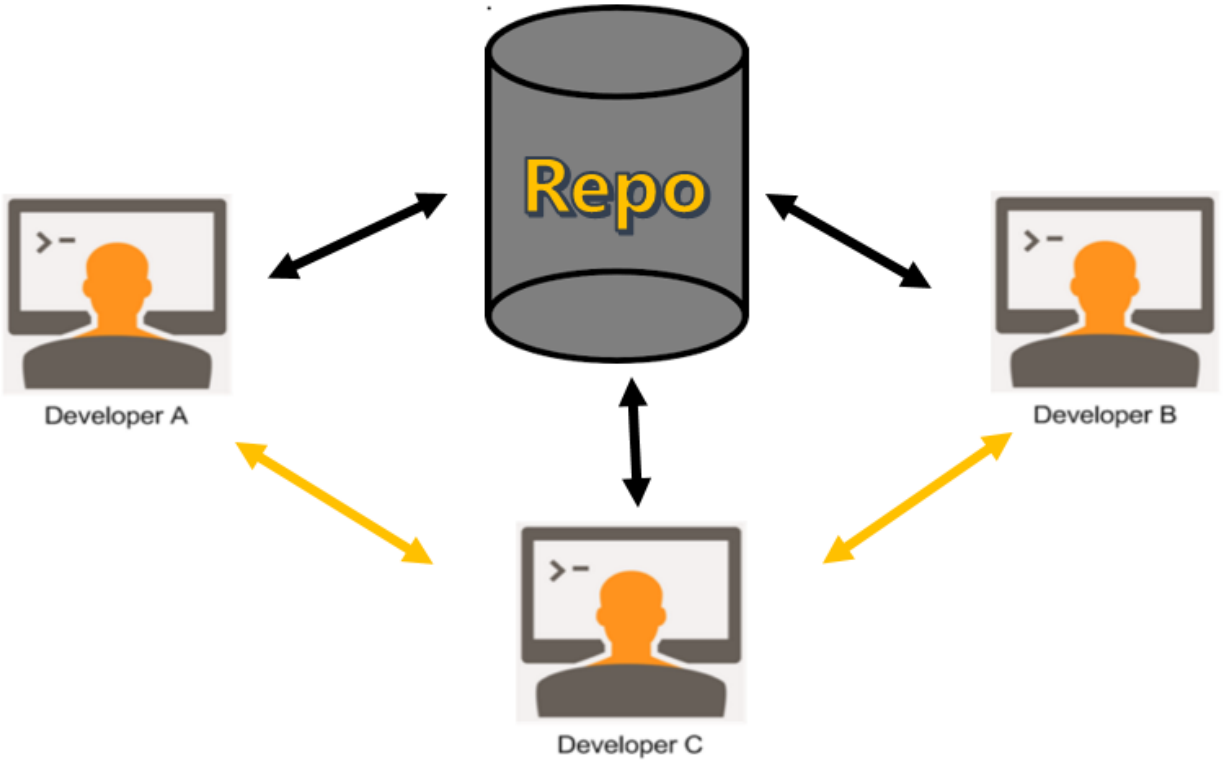


Figure 3.9:

3.3.8 Notebooks: R Markdown

Notebooks: The interactive document for programming in jupyter (julia, python, and r programming).

R Markdown: `A set of code` to convert plain text to HTML, pdf, or latex direct in RStudio.

Chapter 4

Terminology

4.1 Data Cleaning

4.1.1 Messy vs Tidy data

> "Happy families are all alike; every unhappy family is unhappy in its own way."-Leo Tolstoy
>
> "Tidy datasets are all alike, but every messy dataset is messy in its own way."-Hadley Wickham (Chief Data Scientist, Rstudio)

Five Elements of Messy Data:

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

See the journal of the Tidy Data from Journal of Statistical Software:

- http://vita.had.co.nz/papers/tidy-data.pdf
- https://www.jstatsoft.org/index

Three Elements of Tidy Data:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table.

(See Figure 4.1)

This Tidy Data principles follows the same principles as Codd's normalization (focus on a single dataset versus connected datasets like relational databases). More about Codd's normalization process

Note :

For messy data is OpenRefine (formerly Google Refine) a powerful tool for working with. `See the chapter 3.3.2`

Source:

- `R for Data Science [35]`

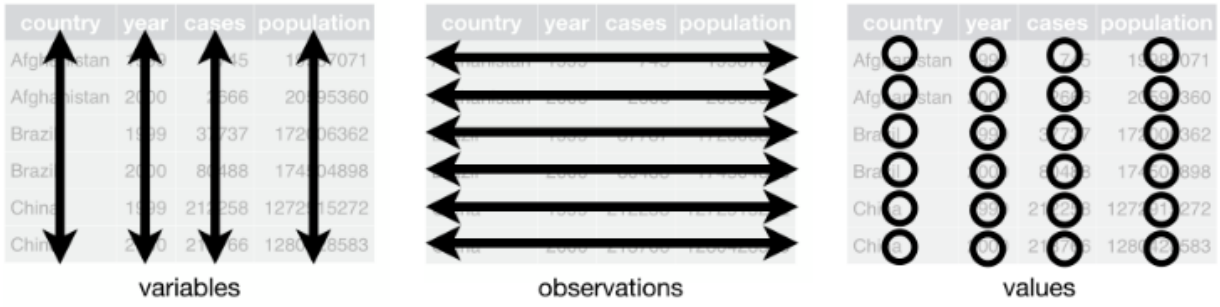


Figure 4.1:

4.1.2 Checklist for Data Cleaning

1-Missing values:

- A placeholder for a datum of which the type is known but its value isn't.
- It is impossible to perform statistical analysis

Use case 1: Identify missing values

1. Statistical analysis error

```
age <- c(23, 16, NA)
mean(age)
```

```
## [1] NA
```

```
## [1] NA
mean(age, na.rm = TRUE)
```

```
## [1] 19.5
```

```
## [1] 19.5
```

2. Identify the NA value

```
complete.cases(age)
```

```
## [1]  TRUE  TRUE FALSE
```

```
#or

is.na(age)
```

```
## [1] FALSE FALSE  TRUE
```

3. Remove NA values (without NA values)

```
na.omit(age)
```

```
## [1] 23 16
## attr(,"na.action")
## [1] 3
## attr(,"class")
## [1] "omit"
```

Use case 2: Recode missing values with mean

1. Test for missing values NA

```
x <- c(1:5, NA, 9:11, NA)
is.na(x)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE
```

```
df <- data.frame(col1 = c(1:3, NA),
                 col2 = c("this", NA,"is", "text"),
                 col3 = c(TRUE, FALSE, TRUE, TRUE),
                 col4 = c(2.5, 4.2, 3.2, NA),
                 stringsAsFactors = FALSE)
is.na(df)
```

```
##       col1  col2  col3  col4
## [1,] FALSE FALSE FALSE FALSE
## [2,] FALSE  TRUE FALSE FALSE
## [3,] FALSE FALSE FALSE FALSE
## [4,]  TRUE FALSE FALSE  TRUE
```

To identify the location of the NA.

```
which(is.na(x))
```

```
## [1]  6 10
```

```
sum(is.na(df))
```

```
## [1] 3
```

```
#or for data frame

colSums(is.na(df))
```

```
## col1 col2 col3 col4
##    1    1    0    1
```

2. Recode missing values NA. Use normal subsetting and assignment operations.

```
# recode missing values with the mean
# vector with missing data
x <- c(1:4, NA, 6:7, NA)
x
```

```
## [1]  1  2  3  4 NA  6  7 NA
```

```
## [1]  1  2  3  4 NA  6  7 NA

x[is.na(x)] <- mean(x, na.rm = TRUE)

round(x, 2)
```

```
## [1] 1.00 2.00 3.00 4.00 3.83 6.00 7.00 3.83
```

```
## [1] 1.00 2.00 3.00 4.00 3.83 6.00 7.00 3.83

# data frame that codes missing values as 99
df <- data.frame(col1 = c(1:3, 99), col2 = c(2.5, 4.2, 99, 3.2))

# change 99s to NAs
df[df == 99] <- NA
df
```

```
##   col1 col2
## 1    1  2.5
## 2    2  4.2
## 3    3   NA
## 4   NA  3.2
```


```
##   col1 col2
## 1    1  2.5
## 2    2  4.2
## 3    3   NA
## 4   NA  3.2
```


Recode missing values in a single data frame.

For example, here we recode the missing value in col4 with the mean value of col4.

```
# data frame with missing data
df <- data.frame(col1 = c(1:3, NA),
                 col2 = c("this", NA,"is", "text"),
                 col3 = c(TRUE, FALSE, TRUE, TRUE),
                 col4 = c(2.5, 4.2, 3.2, NA),
                 stringsAsFactors = FALSE)

df$col4[is.na(df$col4)] <- mean(df$col4, na.rm = TRUE)
df
```

```
##   col1 col2  col3 col4
## 1    1 this  TRUE  2.5
## 2    2 <NA> FALSE  4.2
## 3    3   is  TRUE  3.2
## 4   NA text  TRUE  3.3
```


```
##   col1 col2  col3 col4
## 1    1 this  TRUE  2.5
## 2    2 <NA> FALSE  4.2
## 3    3   is  TRUE  3.2
## 4   NA text  TRUE  3.3
```


3. Exclude missing values NA

```
# A vector with missing values
x <- c(1:4, NA, 6:7, NA)

# including NA values will produce an NA output
mean(x)
```

```
## [1] NA
```


```
## [1] NA
```


```
# excluding NA values will calculate the mathematical operation for all non-missing values
mean(x, na.rm = TRUE)
```

```
## [1] 3.833333
```


```
## [1] 3.833333
```


For complete observations.

```
# data frame with missing values
df <- data.frame(col1 = c(1:3, NA),
                 col2 = c("this", NA,"is", "text"),
                 col3 = c(TRUE, FALSE, TRUE, TRUE),
                 col4 = c(2.5, 4.2, 3.2, NA),
                 stringsAsFactors = FALSE)

df
```

```
##   col1 col2  col3 col4
## 1    1 this  TRUE  2.5
## 2    2 <NA> FALSE  4.2
## 3    3   is  TRUE  3.2
## 4   NA text  TRUE   NA
```


```
##   col1 col2  col3 col4
## 1    1 this  TRUE  2.5
## 2    2 <NA> FALSE  4.2
## 3    3   is  TRUE  3.2
## 4   NA text  TRUE   NA
```


```
complete.cases(df)
```

```
## [1]  TRUE FALSE  TRUE FALSE
```


```
## [1]  TRUE FALSE  TRUE FALSE
```


```
# subset with complete.cases to get complete cases
df[complete.cases(df), ]
```

```
##   col1 col2 col3 col4
## 1    1 this TRUE  2.5
## 3    3   is TRUE  3.2
```


```
##   col1 col2 col3 col4
## 1    1 this TRUE  2.5
## 3    3   is TRUE  3.2
```


```
# or subset with `!` operator to get incomplete cases
df[!complete.cases(df), ]
```

```
##   col1 col2  col3 col4
## 2    2 <NA> FALSE  4.2
## 4   NA text  TRUE   NA
```


```
##   col1 col2  col3 col4
## 2    2 <NA> FALSE  4.2
## 4   NA text  TRUE   NA
```


Alternativ with `omit.na()` function

```
# or use na.omit() to get same as above
na.omit(df)
```

```
##   col1 col2 col3 col4
## 1    1 this TRUE  2.5
## 3    3   is TRUE  3.2
```

```
##   col1 col2 col3 col4
## 1    1 this TRUE  2.5
## 3    3   is TRUE  3.2
```

Use case 3: Exploring missing values with naniar package

naniar package provides principled, tidy ways to summarise, visualise, and manipulate missing data with minimal deviations from the workflows in ggplot2 and tidy data.

http://naniar.njtierney.com/index.html

Use case 4: Missing Value Treatment

Data prep and pattern:

Using BostonHousing dataset in `mlbench` package. The original BostonHousing data doesn't have missing values, but we injected this dataset with mising values.

```
data ("BostonHousing", package="mlbench")
```

```
original <- BostonHousing  # backup original data
```

Introduce missing values randomly `set.seed(100)`

```
BostonHousing[sample(1:nrow(BostonHousing), 40), "rad"] <- NA #40 NA in "rad"
```

```
BostonHousing[sample(1:nrow(BostonHousing), 40), "ptratio"]<-NA #40 NA in "ptratio"
```

```
BostonHousing
```

```
#      crim    zn indus chas    nox    rm   age    dis rad tax ptratio      b lstat medv
# 1  0.00632  18.0  2.31    0 0.5380 6.575  65.2 4.0900   1 296    15.3 396.90  4.98 24.0
# 2  0.02731   0.0  7.07    0 0.4690 6.421  78.9 4.9671   2 242    17.8 396.90  9.14 21.6
# 3  0.02729   0.0  7.07    0 0.4690 7.185  61.1 4.9671   2 242    17.8 392.83  4.03 34.7
# 4  0.03237   0.0  2.18    0 0.4580 6.998  45.8 6.0622   3 222    18.7 394.63  2.94 33.4
# 5  0.06905   0.0  2.18    0 0.4580 7.147  54.2 6.0622   3 222    18.7 396.90  5.33 36.2
# 6  0.02985   0.0  2.18    0 0.4580 6.430  58.7 6.0622   3 222    18.7 394.12  5.21 28.7
# 7  0.08829  12.5  7.87    0 0.5240 6.012  66.6 5.5605   5 311    15.2 395.60 12.43 22.9
# 8  0.14455  12.5  7.87    0 0.5240 6.172  96.1 5.9505   5 311    15.2 396.90 19.15 27.1
# 9  0.21124  12.5  7.87    0 0.5240 5.631 100.0 6.0821   5 311    15.2 386.63 29.93 16.5
# 10 0.17004  12.5  7.87    0 0.5240 6.004  85.9 6.5921   5 311    15.2 386.71 17.10 18.9
# 11 0.22489  12.5  7.87    0 0.5240 6.377  94.3 6.3467   5 311    15.2 392.52 20.45 15.0
# 12 0.11747  12.5  7.87    0 0.5240 6.009  82.9 6.2267   5 311    15.2 396.90 13.27 18.9
# 13 0.09378  12.5  7.87    0 0.5240 5.889  39.0 5.4509   5 311    15.2 390.50 15.71 21.7
```

The missing values have been injected. Identify the 'missings' pattern using ´mice::md.pattern‘.

```
#Pattern of missing values
```

```
#install.packages("mice")
```

```
#library(mice)
```

```
md.pattern(BostonHousing)  # pattern or missing values in data.
```

(See Figure 4.2)

```
# crim zn indus chas nox rm age dis tax b lstat medv rad ptratio
# 431    1  1     1    1   1  1   1   1   1 1     1    1   1       1  0
# 35     1  1     1    1   1  1   1   1   1 1     1    1   1       0  1
# 35     1  1     1    1   1  1   1   1   1 1     1    1   0       1  1
# 5      1  1     1    1   1  1   1   1   1 1     1    1   0       0  2
#        0  0     0    0   0  0   0   0   0 0     0    0  40      40 80
```

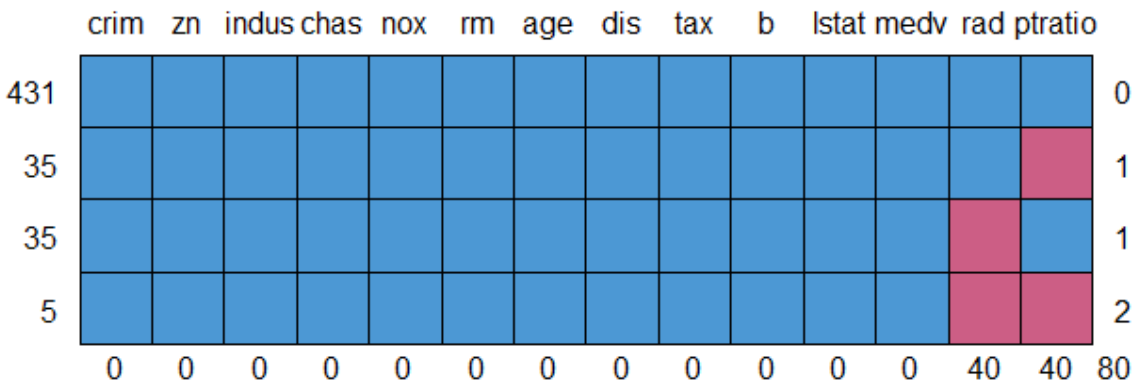

crim zn indus chas nox rm age dis tax b lstat medv rad ptratio
431 0
35 1
35 1
5 2
0 0 0 0 0 0 0 0 0 0 0 0 40 40 80

Figure 4.2:

As we know there are 40 NA values in "rad" and also in "ptratio".

4 ways of dealing with missing values:

1. Deleting the observations.

- Have sufficent data points, so the model doesn't lose power.
- Not to introduce bias (meaning, disproportionate or non-representation of classes).

Using na.action=na.omit

```
cevi<-c(1,2,NA)
cevi
```

```
## [1]  1  2 NA
```

```
cevi_new<-na.omit(cevi)
cevi_new
```

```
## [1] 1 2
## attr(,"na.action")
## [1] 3
## attr(,"class")
## [1] "omit"
```

2. Deleting the variable.

Delete the variable with missing values ("rad" and also "ptratio").

3. Imputation with mean / median / mode.

Replacing the missing values with the mean / median / mode is a crude way of treating missing values. Depending on the context, like if the variation is low or if the variable has low leverage over the response, such a rough approximation is acceptable and could possibly give satisfactory results.

Change the missing values using Hmisc package.

```
install.packages("Hmisc")
```

```
library(Hmisc)
```

```
impute(BostonHousing$ptratio, mean)  # replace with mean
```

```
impute(BostonHousing$rad, mean)  # replace with mean
```

```
#impute(BostonHousing$ptratio, mean)  # replace with mean
 #  [1] 15.3000 17.8000 17.8000 18.7000 18.7000 18.7000 15.2000 15.2000 15.2000 15.2000 15.2000 15.2000
 # [13] 15.2000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000
 # [25] 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 21.0000 19.2000 19.2000
 # [37] 19.2000 19.2000 19.2000 18.3000 18.3000 17.9000 18.4676 17.9000 17.9000 17.9000 17.9000 17.9000



#impute(BostonHousing$rad, mean)  # replace with mean
  #        1        2        3        4        5        6        7        8        9
  # 1.000000 2.000000 2.000000 3.000000 3.000000 3.000000 5.000000 5.000000 5.000000
  #       10       11       12       13       14       15       16       17       18
  # 5.000000 5.000000 5.000000 5.000000 4.000000 4.000000 4.000000 4.000000 4.000000
  #       19       20       21       22       23       24       25       26       27
  # 4.000000 4.000000 4.000000 4.000000 4.000000 4.000000 4.000000 4.000000 4.000000
```

More info about this functions:

https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/regr.eval

2-Special values:

Special values are a values that are not an element of the mathematical set of real numbers. (example: ±Inf and NaN).

A function to check special values in a data.frame:

```
age<-c(21,42,33,18,21,NA,Inf,NaN)
height<-c(111,112,113,114,115,116,117,118)
person<-data.frame(age,height)

is.finite(c(1, Inf, NaN, NA))
```

```
## [1]  TRUE FALSE FALSE FALSE
```

```
is.special <- function(x){
if (is.numeric(x)) !is.finite(x) else is.na(x)
}

sapply(person, is.special)
```

```
##        age height
## [1,] FALSE  FALSE
## [2,] FALSE  FALSE
## [3,] FALSE  FALSE
## [4,] FALSE  FALSE
## [5,] FALSE  FALSE
## [6,]  TRUE  FALSE
## [7,]  TRUE  FALSE
## [8,]  TRUE  FALSE
```

```
person[sapply(person, is.special),]
```

```
##   age height
## 6  NA    116
## 7 Inf    117
## 8 NaN    118
```

3-Duplicate values:

- R base functions:
  - `duplicated()`: for identifying duplicated elements and
  - `unique()`: for extracting unique elements
- `distinct()` function from **dplyr** package to remove duplicate rows in a data frame.

(See Figure 4.3)

Package:

```
install.packages("tidyverse")
```

```
library(tidyverse)
```

Dataset:

```
my_data <- as_tibble(iris) #ribble aren't different from dataframe
```

```
my_data
```

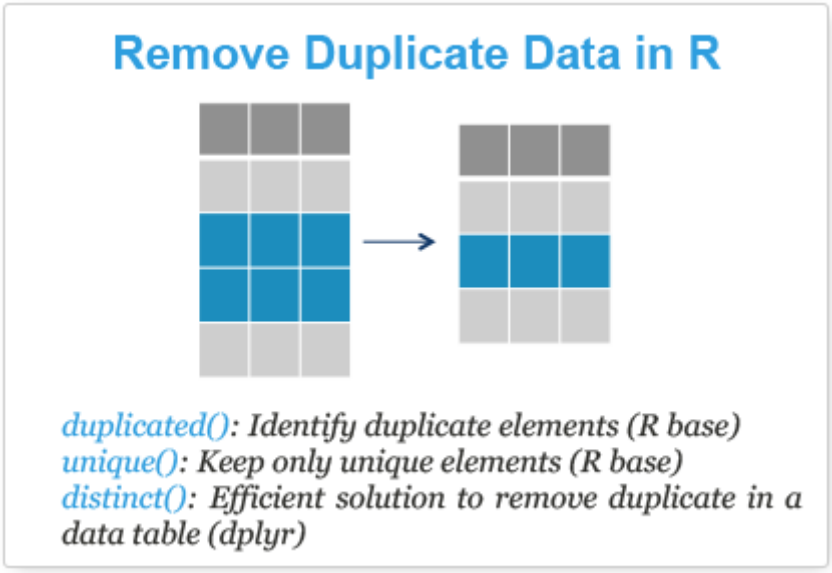


Figure 4.3:

```
# > my_data
# # A tibble: 150 x 5
#    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
#           <dbl>       <dbl>        <dbl>       <dbl> <fct>
#  1          5.1         3.5          1.4         0.2 setosa
#  2          4.9         3            1.4         0.2 setosa
#  3          4.7         3.2          1.3         0.2 setosa
#  4          4.6         3.1          1.5         0.2 setosa
#  5          5           3.6          1.4         0.2 setosa
#  6          5.4         3.9          1.7         0.4 setosa
#  7          4.6         3.4          1.4         0.3 setosa
#  8          5           3.4          1.5         0.2 setosa
#  9          4.4         2.9          1.4         0.2 setosa
# 10          4.9         3.1          1.5         0.1 setosa
# # ... with 140 more rows
```

Remove duplicate rows from a data frame:

Remove duplicates based on Sepal.Width columns.

```
my_data[!duplicated(my_data$Sepal.Width), ]
```

```
#  A tibble: 23 x 5
#    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
#           <dbl>       <dbl>        <dbl>       <dbl> <fct>
#  1          5.1         3.5          1.4         0.2 setosa
#  2          4.9         3            1.4         0.2 setosa
#  3          4.7         3.2          1.3         0.2 setosa
#  4          4.6         3.1          1.5         0.2 setosa
#  5          5           3.6          1.4         0.2 setosa
#  6          5.4         3.9          1.7         0.4 setosa
#  7          4.6         3.4          1.4         0.3 setosa
#  8          4.4         2.9          1.4         0.2 setosa
```

```
#  9          5.4         3.7          1.5         0.2 setosa
# 10          5.8         4            1.2         0.2 setosa
#  ... with 13 more rows
```

Extract unique elements:

```
x <- c(1, 1, 4, 5, 4, 6)
unique(x)
```

```
## [1] 1 4 5 6
```

Remove duplicate rows in a data frame:

```
my_data %>% distinct()
```

```
# # A tibble: 149 x 5
#    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
#           <dbl>       <dbl>        <dbl>       <dbl> <fct>
#  1          5.1         3.5          1.4         0.2 setosa
#  2          4.9         3            1.4         0.2 setosa
#  3          4.7         3.2          1.3         0.2 setosa
#  4          4.6         3.1          1.5         0.2 setosa
#  5          5           3.6          1.4         0.2 setosa
#  6          5.4         3.9          1.7         0.4 setosa
#  7          4.6         3.4          1.4         0.3 setosa
#  8          5           3.4          1.5         0.2 setosa
#  9          4.4         2.9          1.4         0.2 setosa
# 10          4.9         3.1          1.5         0.1 setosa
# # ... with 139 more rows
```

Remove duplicate rows based on certain columns (variables):

```
# Remove duplicated rows based on Sepal.Length
```

```
my_data %>% distinct(Sepal.Length, .keep_all = TRUE)
```

```
# # A tibble: 35 x 5
#    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
#           <dbl>       <dbl>        <dbl>       <dbl> <fct>
#  1          5.1         3.5          1.4         0.2 setosa
#  2          4.9         3            1.4         0.2 setosa
#  3          4.7         3.2          1.3         0.2 setosa
#  4          4.6         3.1          1.5         0.2 setosa
#  5          5           3.6          1.4         0.2 setosa
#  6          5.4         3.9          1.7         0.4 setosa
#  7          4.4         2.9          1.4         0.2 setosa
#  8          4.8         3.4          1.6         0.2 setosa
#  9          4.3         3            1.1         0.1 setosa
# 10          5.8         4            1.2         0.2 setosa
# # ... with 25 more rows
```

```
#Remove duplicated rows based on  Sepal.Length and Petal.Width
```

```
my_data %>% distinct(Sepal.Length, Petal.Width, .keep_all = TRUE)
```

```
#The option .kep_all is used to keep all variables in the data.
```

```
# # A tibble: 110 x 5
#    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
#           <dbl>       <dbl>        <dbl>       <dbl> <fct>
```

```
#  1          5.1         3.5          1.4         0.2 setosa
#  2          4.9         3            1.4         0.2 setosa
#  3          4.7         3.2          1.3         0.2 setosa
#  4          4.6         3.1          1.5         0.2 setosa
#  5          5           3.6          1.4         0.2 setosa
#  6          5.4         3.9          1.7         0.4 setosa
#  7          4.6         3.4          1.4         0.3 setosa
#  8          4.4         2.9          1.4         0.2 setosa
#  9          4.9         3.1          1.5         0.1 setosa
# 10          5.4         3.7          1.5         0.2 setosa
# # ... with 100 more rows
```

Summary:

- Remove duplicate rows based on one or more column values: my_data %>% dplyr::distinct(Sepal.Length)
- R based function: `unique(my_data)`
- Identify duplicate values: `duplicated(my_data)`

4-Outliers:

The simple method for outlier detection is with `dlookr` or `DataExplorer` package. It based on `IQR method`. More info about IQR methode: http://www.mathwords.com/o/outlier.htm

Create an automated report with this function:

- `dlookr`: `diagnose_report(data,output_format ="html")`
- `DataExplorer`: `create_report(data,output_format = html_document(toc = TRUE, toc_depth = 6, theme = "yeti"))`

Sources:

- `R for Data Science [35]`
- `An introduction to data cleaning with R [36]`
- `Identify and Remove Duplicate Data in R [38]`

4.2 EDA: Exploratory Data Analysis

Data quality report: Complete Report about data quality. There are many methods and software doing this step. For example with R Programming.

1. With `dlookr` package from R Programming:

Overview: Diagnose, explore and transform data with `dlookr`. The name dlookr comes from `looking at the data in the data analysis process`.

Features:

- Diagnose data quality.
- Find appropriate scenarios to pursuit the follow-up analysis through data exploration and understanding.
- Derive new variables or perform variable transformations.
- Automatically generate reports for the above three tasks.

Install dlookr:

```
install.packages("dlookr")
```

```
library(dlookr)
```

Usage:

- Data quality diagnosis for data.frame (rows column data), tbl_df, and table of DBMS
- Exploratory Data Analysis for data.frame, tbl_df, and table of DBMS (Database Management System)
- Data Transformation
- Data diagnosis and EDA for table of DBMS (Database Management System)

Data quality diagnosis:

Data: `nycflights13` from the `nycflights13 package`.

General diagnosis of all variables with `diagnose()`:

The variables of the tbl_df object returned by diagnose () are as follows.

- `variables` : variable name
- `types` : the data type of the variable
- `missing_count` : number of missing values
- `missing_percent` : percentage of missing values
- `unique_count` : number of unique values
- `unique_rate` : rate of unique value. unique_count / number of observation
- `Missing Value(NA)` : Variables with very large missing values, i.e. those with a missing_percent close to 100, should be excluded from the analysis.
- `Unique value` : Variables with a unique value (unique_count = 1) are considered to be excluded from data analysis. And if the data type is not numeric (integer, numeric) and the number of unique values is equal to the number of observations (unique_rate = 1), then the variable is likely to be an identifier. Therefore, this variable is also not suitable for the analysis model.

```
diagnose(nycflights13::flights)
```

```
#> # A tibble: 19 x 6
#>    variables   types missing_count missing_percent unique_count unique_rate
#>    <chr>       <chr>         <int>           <dbl>        <int>       <dbl>
#>  1 year        inte…             0           0                1  0.00000297
#>  2 month       inte…             0           0               12  0.0000356
#>  3 day         inte…             0           0               31  0.0000920
#>  4 dep_time    inte…          8255           2.45          1319  0.00392
#>  5 sched_dep_… inte…             0           0             1021  0.00303
#>  6 dep_delay   nume…          8255           2.45           528  0.00157
#>  7 arr_time    inte…          8713           2.59          1412  0.00419
#>  8 sched_arr_… inte…             0           0             1163  0.00345
#>  9 arr_delay   nume…          9430           2.80           578  0.00172
#> 10 carrier     char…             0           0               16  0.0000475
#> 11 flight      inte…             0           0             3844  0.0114
#> 12 tailnum     char…          2512           0.746         4044  0.0120
#> 13 origin      char…             0           0                3  0.00000891
#> 14 dest        char…             0           0              105  0.000312
#> 15 air_time    nume…          9430           2.80           510  0.00151
#> 16 distance    nume…             0           0              214  0.000635
#> 17 hour        nume…             0           0               20  0.0000594
#> 18 minute      nume…             0           0               60  0.000178
#> 19 time_hour   POSI…             0           0             6936  0.0206
```

Diagnosis of numeric variables with `diagnose_numeric()`:

`diagnose_numeric()` diagnoses only numeric(continuous and discrete) variables in a data frame. The variables of the data returned by `diagnose_numeric()` are as follows.

- `min` : minimum value
- `Q1` : 1/4 quartile, 25th percentile
- `mean` : arithmetic mean

- `median` : median, 50th percentile
- `Q3` : 3/4 quartile, 75th percentile
- `max` : maximum value
- `zero` : number of observations with a value of 0
- `minus` : number of observations with negative numbers
- `outlie` : number of outliers

```
diagnose_numeric(nycflights13::flights)
```

```
#> # A tibble: 14 x 10
#>    variables       min    Q1   mean median    Q3   max  zero  minus outlier
#>    <chr>         <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl> <int>  <int>   <int>
#>  1 year           2013  2013 2.01e3   2013  2013  2013     0      0       0
#>  2 month             1     4 6.55e0      7    10    12     0      0       0
#>  3 day               1     8 1.57e1     16    23    31     0      0       0
#>  4 dep_time          1   907 1.35e3   1401  1744  2400     0      0       0
#>  5 sched_dep_ti…   106   906 1.34e3   1359  1729  2359     0      0       0
#>  6 dep_delay       -43    -5 1.26e1     -2    11  1301 16514 183575   43216
#>  7 arr_time          1  1104 1.50e3   1535  1940  2400     0      0       0
#>  8 sched_arr_ti…     1  1124 1.54e3   1556  1945  2359     0      0       0
#>  9 arr_delay       -86   -17 6.90e0     -5    14  1272  5409 188933   27880
#> 10 flight            1   553 1.97e3   1496  3465  8500     0      0       1
#> 11 air_time         20    82 1.51e2    129   192   695     0      0    5448
#> 12 distance         17   502 1.04e3    872  1389  4983     0      0     715
#> 13 hour              1     9 1.32e1     13    17    23     0      0       0
#> 14 minute            0     8 2.62e1     29    44    59 60696      0       0
```

Diagnosis of categorical variables with `diagnose_category()`:

`diagnose_category()` diagnoses the categorical(factor, ordered, character) variables of a data. They are include,

- `variables` : variable names
- `levels`: level names
- `N` : Number of observation
- `freq` : Number of observation at the levles
- `ratio` : Percentage of observation at the levles
- `rank` : Rank of occupancy ratio of levels

```
diagnose_category(nycflights13::flights)
```

```
#> # A tibble: 33 x 6
#>    variables levels      N  freq ratio  rank
#>    <chr>     <chr>   <int> <int> <dbl> <int>
#>  1 carrier   UA     336776 58665 17.4      1
#>  2 carrier   B6     336776 54635 16.2      2
#>  3 carrier   EV     336776 54173 16.1      3
#>  4 carrier   DL     336776 48110 14.3      4
#>  5 carrier   AA     336776 32729  9.72     5
#>  6 carrier   MQ     336776 26397  7.84     6
#>  7 carrier   US     336776 20536  6.10     7
#>  8 carrier   9E     336776 18460  5.48     8
#>  9 carrier   WN     336776 12275  3.64     9
#> 10 carrier   VX     336776  5162  1.53    10
#> # … with 23 more rows
```

Diagnosing outliers with `diagnose_outlier()`:

`diagnose_outlier()` diagnoses the outliers of the numeric (continuous and discrete) variables of the data frame (row and column data style, it include:

- outliers_cnt : Count of outliers
- outliers_ratio : Percent of outliers
- outliers_mean : Arithmetic Average of outliers
- with_mean : Arithmetic Average of with outliers
- without_mean : Arithmetic Average of without outliers

```
diagnose_outlier(nycflights13::flights)
```

```
#> # A tibble: 14 x 6
#>    variables outliers_cnt outliers_ratio outliers_mean with_mean
#>    <chr>            <int>          <dbl>         <dbl>     <dbl>
#>  1 year                 0       0                NaN     2013
#>  2 month                0       0                NaN        6.55
#>  3 day                  0       0                NaN       15.7
#>  4 dep_time             0       0                NaN     1349.
#>  5 sched_de…            0       0                NaN     1344.
#>  6 dep_delay        43216      12.8               93.1     12.6
#>  7 arr_time             0       0                NaN     1502.
#>  8 sched_ar…            0       0                NaN     1536.
#>  9 arr_delay        27880       8.28             121.       6.90
#> 10 flight               1       0.000297        8500     1972.
#> 11 air_time          5448       1.62             400.      151.
#> 12 distance           715       0.212           4955.     1040.
#> 13 hour                0       0                NaN       13.2
#> 14 minute              0       0                NaN       26.2
#> # … with 1 more variable: without_mean <dbl>
```

Visualization of outliers using `plot_outlier()`:

The plot derived from outlier data diagnosis is as follows.

- With outliers box plot
- Without outliers box plot
- With outliers histogram
- Without outliers histogram

```
plot_outlier(nycflights13::flights)
```

(See Figure 4.4)

Exploratory Data Analysis:

Datset: The data come from `ISLR package` (Carseats). The data has variables are as follows.

- Sales: Unit sales (in thousands) at each location
- CompPrice: Price charged by competitor at each location
- Income: Community income level (in thousands of dollars)
- Advertising: Local advertising budget for company at each location (in thousands of dollars)
- Population: Population size in region (in thousands)
- Price: Price company charges for car seats at each site
- ShelveLoc: A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site
- Age: Average age of the local population
- Education: Education level at each location
- Urban: A factor with levels No and Yes to indicate whether the store is in an urban or rural location
- US: A factor with levels No and Yes to indicate whether the store is in the US or not

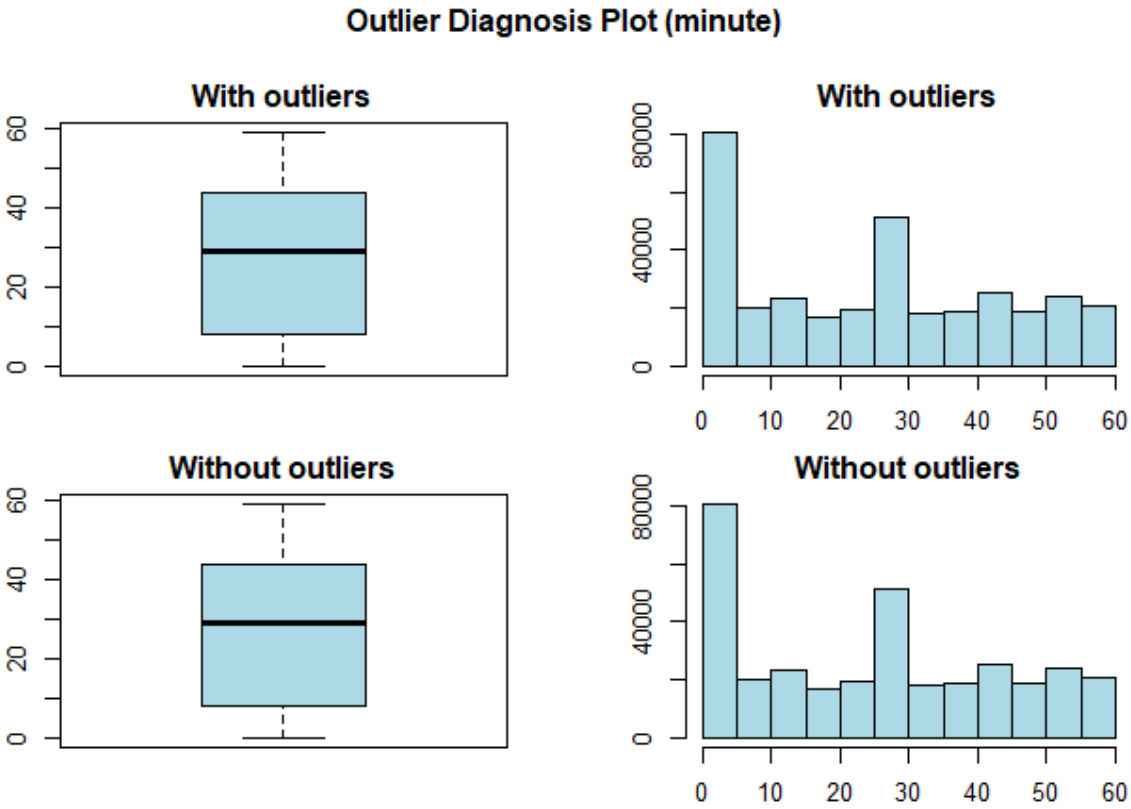

Outlier Diagnosis Plot (minute)
With outliers
With outliers
Without outliers
Without outliers
0
20
40
60
0
40000
80000
0
10
20
30
40
50
60

Figure 4.4:

Univariate data EDA: Calculating descriptive statistics using `describe()`. The 'describe() function has the variable are as follows.

- `n` : number of observations excluding missing values
- `na` : number of missing values
- `mean` : arithmetic average
- `sd` : standard devation
- `se_mean` : standrd error mean. sd/sqrt(n)
- `IQR` : interquartile range (Q3-Q1)
- `skewness` : skewness
- `kurtosis` : kurtosis
- `p25` : Q1. 25% percentile
- `p50` : median. 50% percentile
- `p75` : Q3. 75% percentile
- `p01`, `p05`, `p10`, `p20`, `p30` : 1%, 5%, 20%, 30% percentiles
- `p40`, `p60`, `p70`, `p80` : 40%, 60%, 70%, 80% percentiles
- `p90`, `p95`, `p99`, `p100` : 90%, 95%, 99%, 100% percentiles
- `skewness`: The measure of the asymmetry of the probability distribution. The left-skewed distribution data, that is, the variables with large positive skewness should consider the log or sqrt transformations to follow the normal distribution. The variables Advertising seem to need to consider variable transformations.
- `mean` and `sd`, `se_mean`: Arithmetic mean(mean), the standard deviation(sd), and standard error of the mean(se_mean).

```
describe(ISLR::Carseats)
```

```
#> # A tibble: 8 x 26
#>   variable     n    na   mean     sd se_mean    IQR skewness kurtosis   p00
#>   <chr>    <dbl> <dbl>  <dbl>  <dbl>   <dbl>  <dbl>    <dbl>    <dbl> <dbl>
#> 1 Sales      400     0   7.50   2.82   0.141   3.93   0.186   -0.0809     0
#> 2 CompPri…   400     0 125.    15.3    0.767  20     -0.0428   0.0417    77
#> 3 Income     380    20  68.9   28.1    1.44   48.2    0.0449  -1.09      21
#> 4 Adverti…   400     0   6.64   6.65   0.333  12      0.640   -0.545      0
#> 5 Populat…   400     0 265.   147.     7.37  260.    -0.0512  -1.20      10
#> 6 Price      400     0 116.    23.7    1.18   31     -0.125    0.452     24
#> 7 Age        400     0  53.3   16.2    0.810  26.2   -0.0772  -1.13      25
#> 8 Educati…   400     0  13.9    2.62   0.131   4      0.0440  -1.30      10
#> # … with 16 more variables: p01 <dbl>, p05 <dbl>, p10 <dbl>, p20 <dbl>,
#> #   p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>, p70 <dbl>,
#> #   p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>
```

Test of normality on numeric variables using `normality()`:

The normality test on numerical data. There are many method to indentify normality (ex: graphical test, Kolmogorov–Smirnov test, Shapiro–Wilk test, and Pearson's chi-squared test). In this `normality()` function Shapiro-Wilk normality test is performed.

The variables of data returned by `normality()` are as follows.

- `statistic` : Statistics of the Shapiro-Wilk test
- `p_value` : p-value of the Shapiro-Wilk test
- `sample` : Number of sample observations performed Shapiro-Wilk test

```
# vars        statistic  p_value sample
#   <chr>           <dbl>    <dbl>  <dbl>
# 1 Sales           0.995 2.54e- 1    400
# 2 CompPrice       0.998 9.77e- 1    400
# 3 Income          0.961 8.40e- 9    400
```

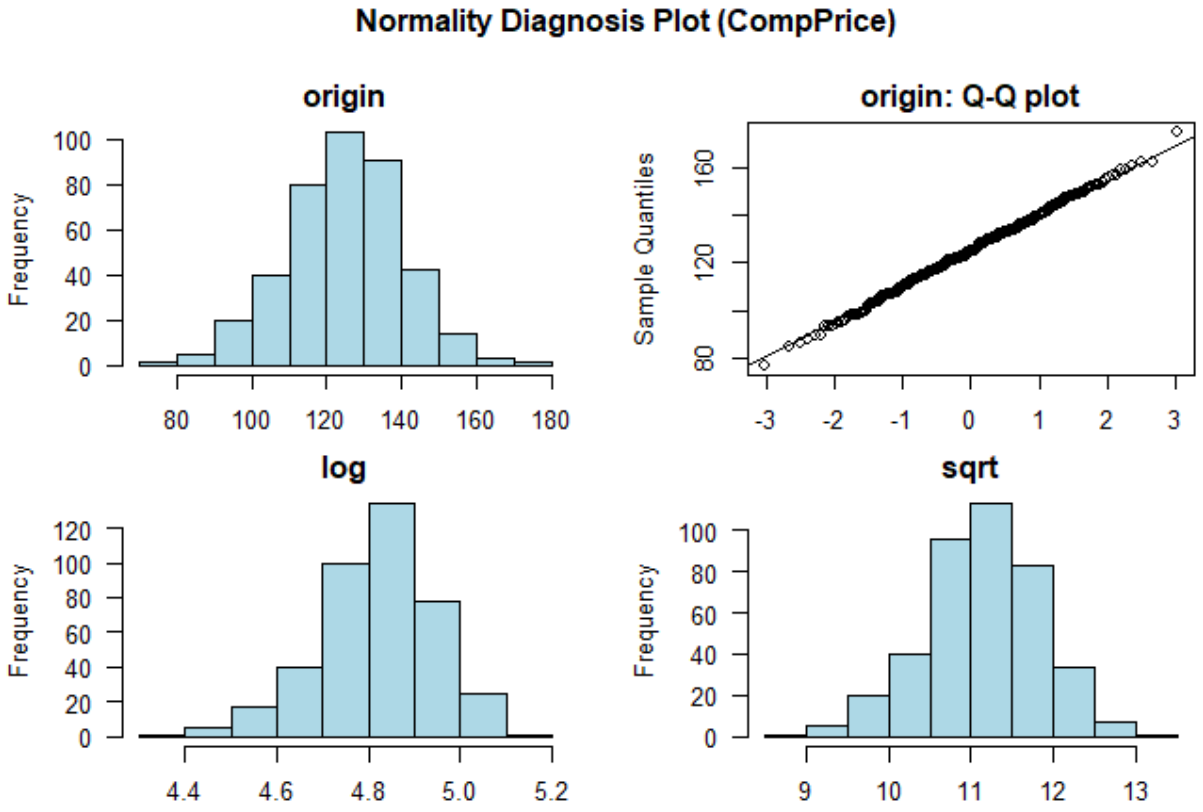


Figure 4.5:

```
# 4 Advertising     0.874 1.49e-17    400
# 5 Population      0.952 4.08e-10    400
# 6 Price           0.996 3.90e- 1    400
# 7 Age             0.957 1.86e- 9    400
# 8 Education       0.924 2.43e-13    400
```

Normalization visualization of numerical variables using `plot_normality()`:

`plot_normality()` visualizes the normality of numeric data, it include:

- `Histogram of original data`
- `Q-Q plot of original data`
- `Histogram of log transformed data`
- `Histogram of square root transformed data`

```
plot_normality(ISLR::Carseats,Sales, CompPrice)
```

(See Figure 4.5)

Bivariate data EDA:

Calculation of correlation coefficient using `correlate()` function.

```
correlate(carseats)
```

```
#> # A tibble: 56 x 3
#>    var1        var2      coef_corr
#>    <fct>       <fct>         <dbl>
```

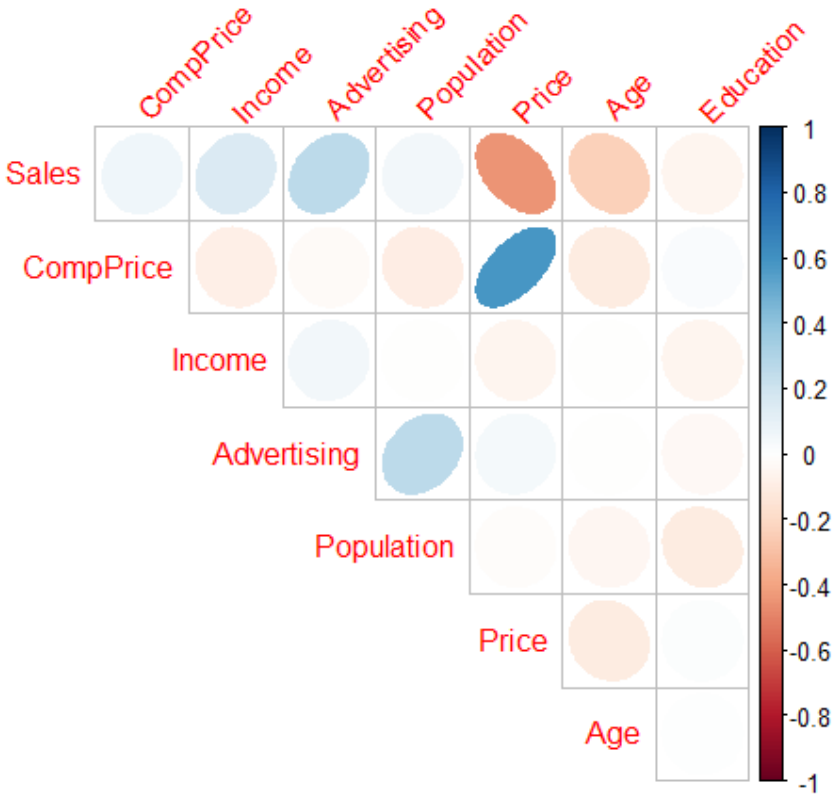


Figure 4.6:

```
#>  1 CompPrice   Sales        0.0641
#>  2 Income      Sales        0.151
#>  3 Advertising Sales        0.270
#>  4 Population  Sales        0.0505
#>  5 Price       Sales       -0.445
#>  6 Age         Sales       -0.232
#>  7 Education   Sales       -0.0520
#>  8 Sales       CompPrice    0.0641
#>  9 Income      CompPrice   -0.0761
#> 10 Advertising CompPrice   -0.0242
#> # … with 46 more rows
```

Visualization of the correlation matrix using `plot_correlate()`:

(See Figure 4.6)

```
plot_correlate(ISLR::Carseats)
```

Note: The diagnose_report() report the information for diagnosing the quality of the data as pdf or html.

Example: `diagnose_report(ISLR::Carseats,output_format ="html")`. It created an automated html EDA report.

Sources:

- https://github.com/choonghyunryu/dlookr &

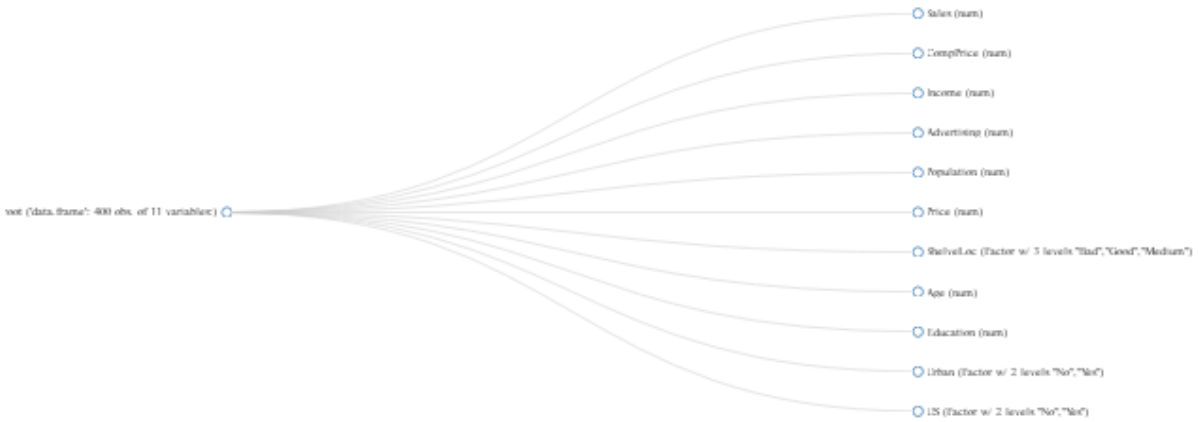


Figure 4.7:

- https://www.rdocumentation.org/packages/dlookr/versions/0.3.9

2. With `DataExplorer package` from R Programming:

Another package from R Programming. We use the same dataset from `ISLR package`: Careseats.

Variables: `plot_str(ISLR::Carseats)`

(See Figure 4.7)

Missing Values: `plot_missing(ISLR::Carseats)`

(See Figure 4.8)

Continuous Variables: `plot_histogram(ISLR::Carseats)` or `plot_density(ISLR::Carseats)`

Multivariate Analysis: `plot_correlation(ISLR::Carseats)`

Categorical Variables-Barplots: `plot_bar(ISLR::Carseats)`

Automated report: `create_report(ISLR::Carseats)`

For more info:

- https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html
- https://www.rdocumentation.org/packages/DataExplorer/versions/0.8.0

Sources:

- `An introduction to data cleaning with R` [36]
- `Identify and Remove Duplicate Data in R` [38]

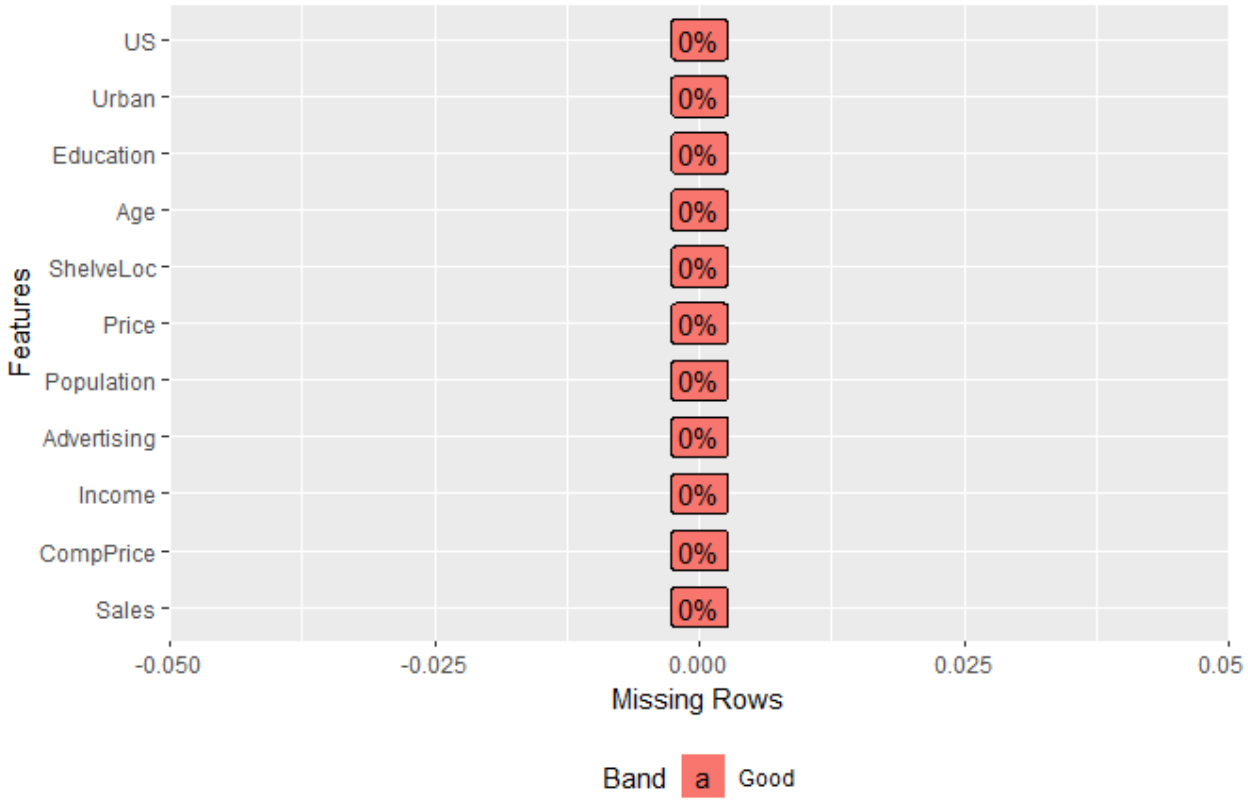

US
Urban
Education
Age
ShelveLoc
Price
Population
Advertising
Income
CompPrice
Sales
0%
Features
-0.050
-0.025
0.000
0.025
0.05
Missing Rows
Band
a
Good

Figure 4.8:

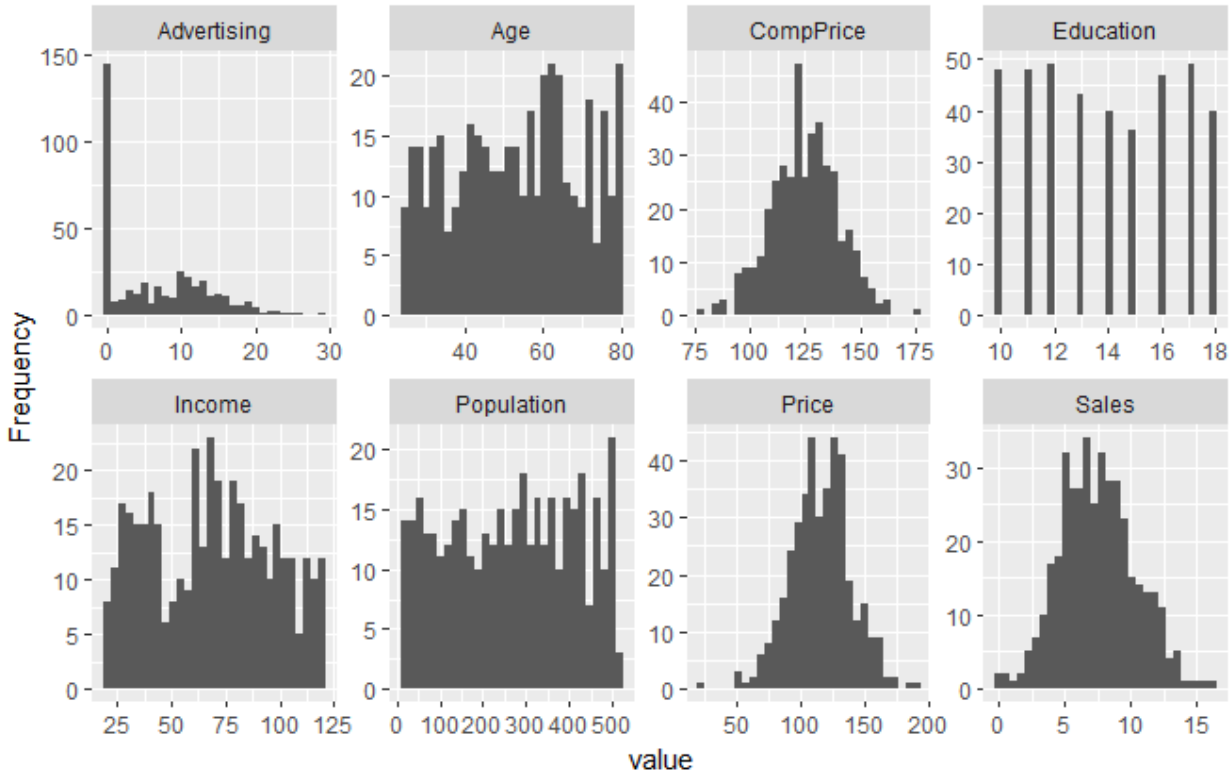

Advertising
Age
CompPrice
Education
Income
Population
Price
Sales
Frequency
value

Figure 4.9:

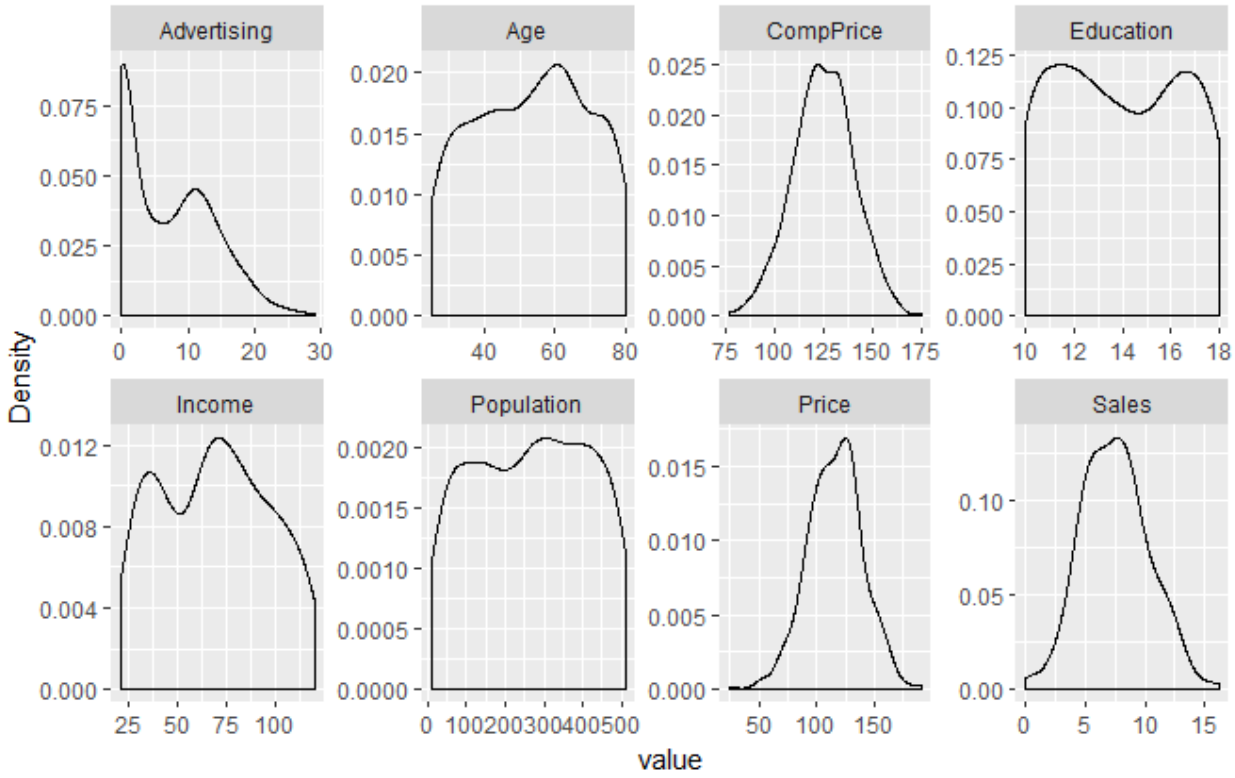

Advertising
Age
CompPrice
Education
Income
Population
Price
Sales
Density
value

Figure 4.10:

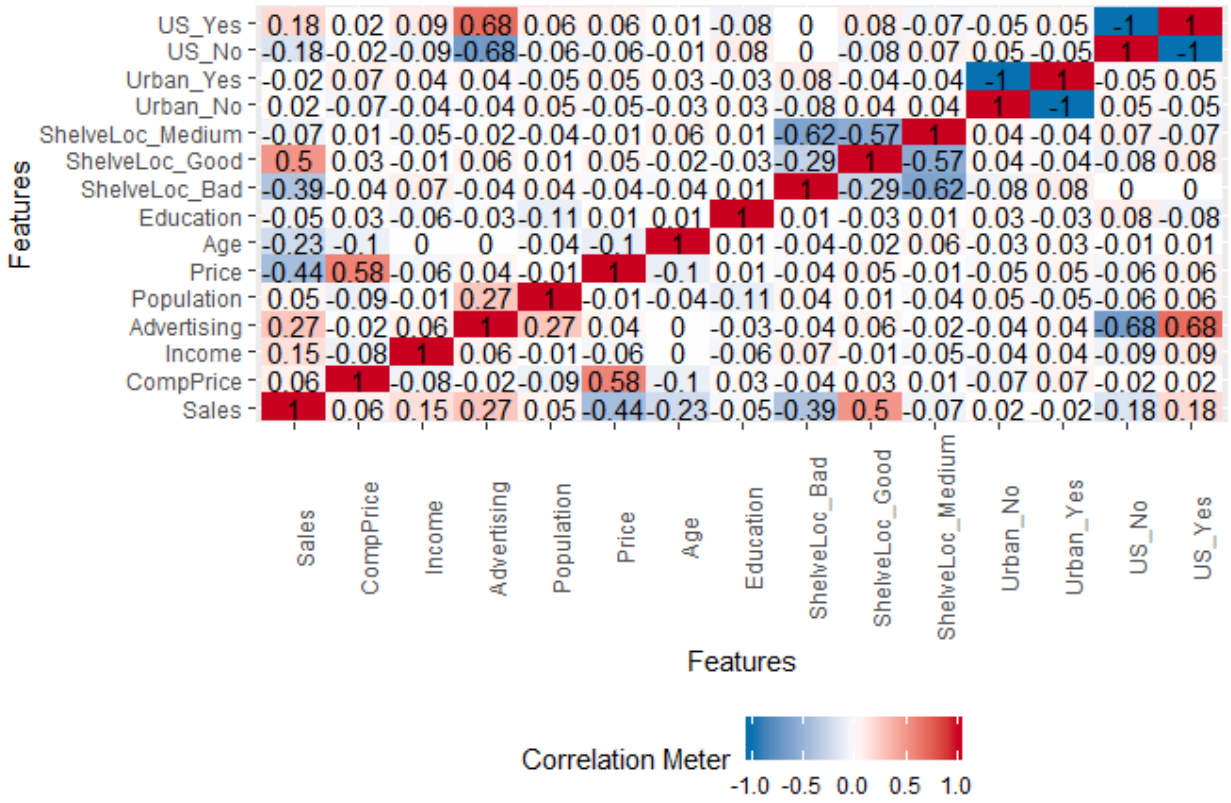

Features
US_Yes 0.18 0.02 0.09 0.68 0.06 0.06 0.01 -0.08 0 0.08 -0.07 -0.05 0.05 -1 1
US_No -0.18 -0.02 -0.09 -0.68 -0.06 -0.06 -0.01 0.08 0 -0.08 0.07 0.05 -0.05 1 -1
Urban_Yes -0.02 0.07 0.04 0.04 -0.05 0.05 0.03 -0.03 0.08 -0.04 -0.04 -1 1 -0.05 0.05
Urban_No 0.02 -0.07 -0.04 -0.04 0.05 -0.05 -0.03 0.03 -0.08 0.04 0.04 1 -1 0.05 -0.05
ShelveLoc_Medium -0.07 0.01 -0.05 -0.02 -0.04 -0.01 0.06 0.01 -0.62 -0.57 1 0.04 -0.04 0.07 -0.07
ShelveLoc_Good 0.5 0.03 -0.01 0.06 0.01 0.05 -0.02 -0.03 -0.29 1 -0.57 0.04 -0.04 -0.08 0.08
ShelveLoc_Bad -0.39 -0.04 0.07 -0.04 0.04 -0.04 -0.04 0.01 1 -0.29 -0.62 -0.08 0.08 0 0
Education -0.05 0.03 -0.06 -0.03 -0.11 0.01 0.01 1 0.01 -0.03 0.01 0.03 -0.03 0.08 -0.08
Age -0.23 -0.1 0 0 -0.04 -0.1 1 0.01 -0.04 -0.02 0.06 -0.03 0.03 -0.01 0.01
Price -0.44 0.58 -0.06 0.04 -0.01 1 -0.1 0.01 -0.04 0.05 -0.01 -0.05 0.05 -0.06 0.06
Population 0.05 -0.09 -0.01 0.27 1 -0.01 -0.04 -0.11 0.04 0.01 -0.04 0.05 -0.05 -0.06 0.06
Advertising 0.27 -0.02 0.06 1 0.27 0.04 0 -0.03 -0.04 0.06 -0.02 -0.04 0.04 -0.68 0.68
Income 0.15 -0.08 1 0.06 -0.01 -0.06 0 -0.06 0.07 -0.01 -0.05 -0.04 0.04 -0.09 0.09
CompPrice 0.06 1 -0.08 -0.02 -0.09 0.58 -0.1 0.03 -0.04 0.03 0.01 -0.07 0.07 -0.02 0.02
Sales 1 0.06 0.15 0.27 0.05 -0.44 -0.23 -0.05 -0.39 0.5 -0.07 0.02 -0.02 -0.18 0.18
Sales
CompPrice
Income
Advertising
Population
Price
Age
Education
ShelveLoc_Bad
ShelveLoc_Good
ShelveLoc_Medium
Urban_No
Urban_Yes
US_No
US_Yes
Features
Correlation Meter
-1.0 -0.5 0.0 0.5 1.0

Figure 4.11:

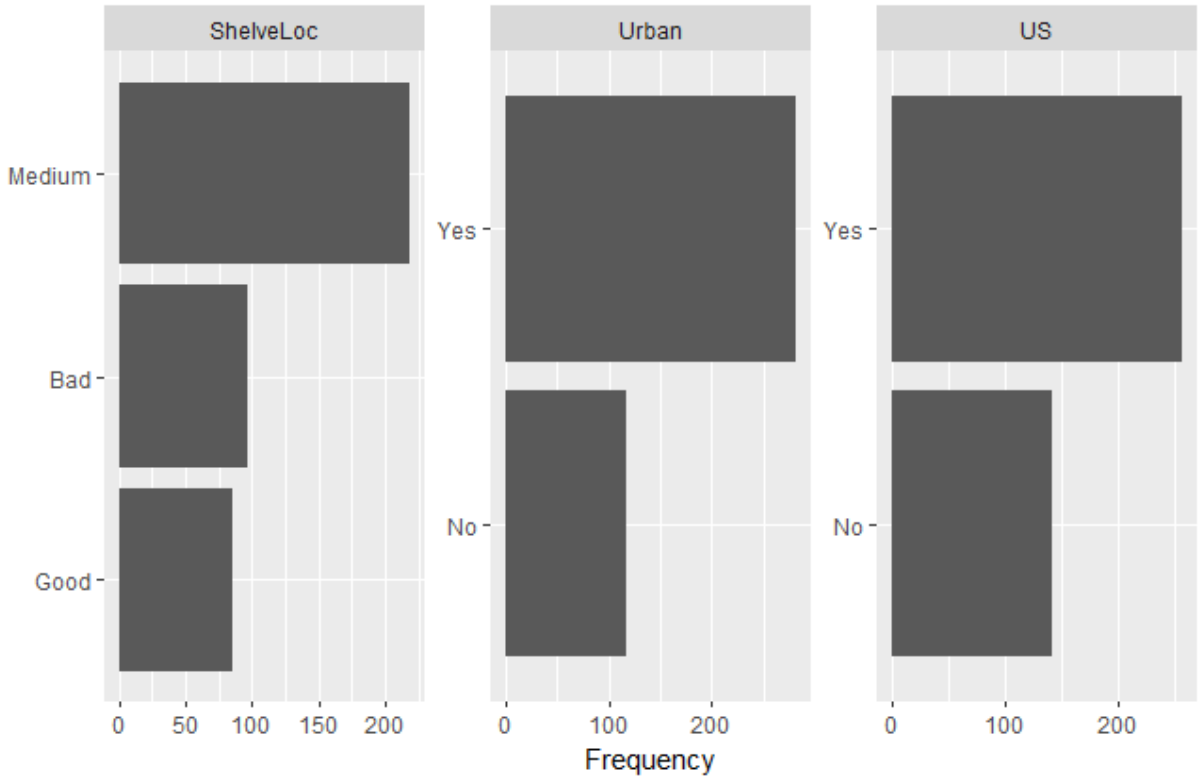

ShelveLoc
Urban
US
Medium
Bad
Good
Yes
No
Yes
No
0
50
100
150
200
0
100
200
0
100
200
Frequency

Figure 4.12:

4.3 DAX: Data Analysis Expressions

Note: I am not writing detail definition or technical view of this term, because this term is really width functions and it needs time to understand the logic.

Key points of DAX:

- Native formula from Microsoft technology
- DAX is not a programming language
- DAX is a library of functions
- It used for Microsoft Excel, Power Pivot, SQL Server Analysis Services (SSAS) , and Microsoft Power BI

More info about DAX: https://docs.microsoft.com/en-us/dax/data-analysis-expressions-dax-reference

Chapter 5

KPIs for big data

What is KPI?

KPI (Key Performance Indicators) are the response to company fear of big data, ugly spreadsheets and uncertain applications with unstructured data. The idea of KPI is that company presenting big data easily and also using business-relevant point of view.

Measuring the right KPI is vital to the health and success of the company. Here are 5 reasons why You need KPIs.

1. Monitor company health
2. Measuring progress over time
3. To stay on track
4. Solving problems and getting more opportunities
5. Analyzing patterns over time

Examples of KPIs are:

- Net Profit Margin
- ROI (Return of Investment)
- Operating Cash Flow (OCF)
- Sales Growth
- BEP (Break Event Point)
- Cash Ratio

Types of KPI

It depends on an organization's structures and goals, the company can track more than point of view. The right KPIs right is crucial for getting actionable and insightful information about the company's performance and situation.

Each department of the company have different KPIs. They have different tasks and goals.

There are five main types of KPIs:

1. Business KPI
2. Financial KPI
3. Sales KPI
4. Marketing KPI
5. Project Management KPI
6. Business KPI

This KPI help to measure long-term business goals. The companies are able to find better insight.

Examples of popular business KPIs are:

- Revenue Growth Rate
- Churn Rate
- Acquisition Rate
- Relative Market Share
- Return on Equity

2. Financial KPI

This KPI provides an assessment of business performance. It used by an organization's leader and finance department to generating better revenue and profits.

Examples of popular financial KPIs are:

- MRR (Monthly Recurring Revenue)
- Net profit margin
- Operating cash flow (OCF)
- Working Capital
- Current Ratio
- Budget Variance

3. Sales KPI

Sales KPIs are measurable values that indicate the performance of various sales processes, used by the sales team to monitor the achievement of their key objectives and goals. Sales metrics help to keep and attain sustainable sales.

Example of Popular sales KPI are:

- Monthly New Leads/Sales
- Lead-to-Customer Conversion Rate
- Cost per Acquisition
- Sales Qualified Leads (SQL)
- Customer Lifetime Value (LTV)

4. Marketing KPI

To help the marketing department to monitor all marketing channels. This KPI shows the marketing teams how to acquiring new leads.

Examples of popular marketing KPI are:

- Website Traffic Per Source
- Cost Per Acquisition CPA
- Marketing Qualified Leads (MQL)
- Net Promoter Score
- Conversion Rate

5. Project Management KPI

To monitor project progress. The Company use project KPI to identify important deadlines.

Popular Project management KPI are:

- Planned Value (PV)
- Actual Cost (AC)
- Earned Value (EV)
- Cost Variance (CV)
- Schedule Variance (SV)

Before we talk about the use case in Deutsche Bahn. We need to understand the types and classification of KPI terms based on perspective, industry, and etc.

`High-level KPI`: For demonstrating the company's overall performance. It includes Annual Growth, Annual Recurring Revenue (ARR), and Relative Market Share.

Single individuals have no impact on high-level KPI. This KPI is the result of teamwork across all multiple departments and subsidiary company.

`Low-level KPI`: For demonstrating single individuals performance. To identify the day-to-day work of a single employee.

Five essential questions for creating a KPI

1. What are the business results (goals)?
2. How can the KPI values be improved by taking action?
3. Do we have all the relevant data?
4. Who is going to use the KPI?
5. How to visualize specific KPI (graphs, metrics, diagrams, etc.)?

How to choose the right KPI

The most important thing is to define business goals. Try to focus on a few key metrics. It should meet the SMART criteria:

SMART KPIs are:

S pecific

M easurable

A ttainable

R elevant

T ime-Bound

Sources:

- `5 Reasons Why You Need The Right KPIs [38]`
- `What Is a KPI? (Complete Guide) [39]`

5.1 Financial Perspective

Based on the size, age, and industry, each and every company needs to be aware of their financial progress. The fastest and simple way to keep track of a company financial situation is to set up a KPI that displays financial point of view. Let's started with the most widely used financial metrics.

Firstly, we need to understand the concept of a balance sheet. Below is the example of a complete balance sheet. (See Figure 5.1)

Example Company
Balance Sheet
December 31, 2018

ASSETS		LIABILITIES	
Current assets		Current liabilities	
Cash	$ 2,100	Notes payable	$ 5,000
Petty cash	100	Accounts payable	35,900
Temporary investments	10,000	Wages payable	8,500
Accounts receivable - net	40,500	Interest payable	2,900
Inventory	31,000	Taxes payable	6,100
Supplies	3,800	Warranty liability	1,100
Prepaid insurance	1,500	Unearned revenues	1,500
Total current assets	89,000	Total current liabilities	61,000
Investments	36,000	Long-term liabilities	
		Notes payable	20,000
Property, plant & equipment		Bonds payable	400,000
Land	5,500	Total long-term liabilities	420,000
Land improvements	6,500		
Buildings	180,000		
Equipment	201,000	Total liabilities	481,000
Less: accum depreciation	(56,000)		
Prop, plant & equip - net	337,000		
Intangible assets		STOCKHOLDERS' EQUITY	
Goodwill	105,000	Common stock	110,000
Trade names	200,000	Retained earnings	220,000
Total intangible assets	305,000	Accum other comprehensive income	9,000
		Less: Treasury stock	(50,000)
Other assets	3,000	Total stockholders' equity	289,000
Total assets	$ 770,000	Total liabilities & stockholders' equity	$ 770,000

The notes to the sample balance sheet have been omitted.

Figure 5.1:

5.1.1 Operating Cash Flow (OCF)

How much money generated by the regular operating activities of a business in a specific time period.

`Short form formula`: Operating Cash Flow = Net Income + Non-Cash Expenses – Increase in Working Capital

`Long form formula` : Operating Cash Flow = Net Income + Depreciation + Stock Based Compensation + Deferred Tax + Other Non Cash Items – Increase of Accounts Receivable – Increase of Inventory + Increase of Accounts Payable + Increase of Accrued Expenses + Increase of Deferred Revenue

Figure 5.2 is an example of operating cash flow (OCF) using Amazon's 2017 annual report.

5.1.2 Current Ratio

The ability to pay all the financial obligations in one year. A good Current Ratio is between 1.5 and 3

`Current Ratio` = Current Assets / Current Liabilities

Example of the Current Ratio Formula:

A business has:

- Cash = $20 million
- Marketable securities = $25 million
- Inventory = $30 million
- Short-term Debt = $20 million
- Accounts Payables = $20 million

then,

- Current assets = 20 + 25 + 30 = 75 million
- Current liabilities = 20 + 20 = 40 million
- Current ratio = 75 million / 40 million = 1.875

The company can easily settle each account payable 1.875x.

5.1.3 Ratio / Acid Test / Quick Ratio

How much the short-term assets to cover its near-future liabilities. The quick ratio deleted liquid assets such as inventories.

`Quick Ratio`: (Current assets – Inventory) / Current Liabilities

Example:

Cash	$70	Accounts Payable	$40
Marketable Securities	$20	Expenses	$30
Account Receivables	$50	Notes Payable	$15
Inventory	$60	Long Term Debt	$20
Total Current Assets	$200	Total Current Liabilities	$105

The Company quick ratio as follows:

($70,000 + $20,000 + $50,000) / $105,000 = 1.27

It means that for every one dollar money of current liabilities, the company has $1.70 liquid assets to cover those liabilities.

AMAZON.COM, INC.

CONSOLIDATED STATEMENTS OF CASH FLOWS

(in millions)

	Year Ended December 31,		
	2015	2016	2017
CASH AND CASH EQUIVALENTS, BEGINNING OF PERIOD	$ 14,557	$ 15,890	$ 19,334
OPERATING ACTIVITIES:			
Net income	596	2,371	3,033
Adjustments to reconcile net income to net cash from operating activities:			
Depreciation of property and equipment, including internal-use software and website development, and other amortization, including capitalized content costs	6,281	8,116	11,478
Stock-based compensation	2,119	2,975	4,215
Other operating expense, net	155	160	202
Other expense (income), net	250	(20)	(292)
Deferred income taxes	81	(246)	(29)
Changes in operating assets and liabilities:			
Inventories	(2,187)	(1,426)	(3,583)
Accounts receivable, net and other	(1,755)	(3,367)	(4,786)
Accounts payable	4,294	5,030	7,175
Accrued expenses and other	913	1,724	283
Unearned revenue	1,292	1,955	738
Net cash provided by (used in) operating activities	12,039	17,272	18,434
INVESTING ACTIVITIES:			
Purchases of property and equipment, including internal-use software and website development	(5,387)	(7,804)	(11,955)
Proceeds from property and equipment incentives	798	1,067	1,897
Acquisitions, net of cash acquired, and other	(795)	(116)	(13,972)
Sales and maturities of marketable securities	3,025	4,733	9,988
Purchases of marketable securities	(4,091)	(7,756)	(13,777)
Net cash provided by (used in) investing activities	(6,450)	(9,876)	(27,819)
FINANCING ACTIVITIES:			
Proceeds from long-term debt and other	353	621	16,231
Repayments of long-term debt and other	(1,652)	(354)	(1,372)
Principal repayments of capital lease obligations	(2,462)	(3,860)	(4,799)
Principal repayments of finance lease obligations	(121)	(147)	(200)
Net cash provided by (used in) financing activities	(3,882)	(3,740)	9,860
Foreign currency effect on cash and cash equivalents	(374)	(212)	713
Net increase (decrease) in cash and cash equivalents	1,333	3,444	1,188
CASH AND CASH EQUIVALENTS, END OF PERIOD	$ 15,890	$ 19,334	$ 20,522

Figure 5.2:

5.1.4 Burn Rate

Burn rate is the amount of time it will take a company to exhaust the capital. Burn Rate is also in terms of cash burned per month, year, or quarterly.

`Burn Rate`= Cash / Expenses

- Gross burn: it includes expenses each month.
- Net burn: it includes loses each month.

Example:

A company spends $10,000 monthly on office space, $20,000 on monthly server costs and $30,000 on salaries and wages for its engineers, its gross burn rate would be $60,000.

However, if the company was already producing revenue, its net burn would be different. Even if the company operates at a loss, with revenues of $40,000 a month and costs of goods sold (COGS) of $20,000, it would still work to reduce its overall burn. In this scenario, the company's net burn would be $20,000, derived as: $40,000 - $20,000 - $30,000 = $10,000.

5.1.5 Net Profit Margin

Indicates how efficient a company at generating profit compared to its revenue.

`Net profit margin` = net profit / revenue

Check this example:

Company ABC and DEF both operate in the same industry. Which company has a higher net profit margin?

Company ABC	Income Statement	Company DEF	Income Statement
Revenue	$200	Revenue	$325
Cost of Goods Sold	$20 ___-	Cost of Goods Sold	$35 ___-
Gross Profit	$80	Gross Profit	$190
Operating Expenses	$20 ___-	Operating Expenses	$40 ___-
Operating Profit	$60	Operating Profit	$150
Interest Expense	$5 ___-	Interest Expense	$10 ___-
Earnings Before Taxes	$55	Earnings Before Taxes	$140
Tax Expense	$25 ___-	Tax Expense	$60 ___-
Net Income	$130	Net Income	$180

Step 1 : Write out the formula

Net Profit Margin = Net Profit/Revenue

Step 2 : Calculate the net profit margin

Company ABC:

Net Profit Margin = Net Profit/Revenue = $130/$200 = 65%

Company DEF:

Net Profit Margin = Net Profit/Revenue = $180/$325 = 55.4%

Company ABC has a higher net profit margin than a company DEF.

5.1.6 Working Capital

The difference between current assets (like cash and goods) and current liabilities (like debts or obligations).

`Working Capital` = Current Assets - Current Liabilities

Example:

Cash	$120	Accounts Payable	$60
Marketable Securities	$20	Expenses	$40
Account Receivables	$80	Notes Payable	$10
Inventory	$100	Long Term Debt	$20
Total Current Assets	$320	Total Current Liabilities	$130

Using the formula,the working capital ist:

$320 - $130,000 = $190

5.1.7 Current Accounts Receivable

How much money owed by its debtors. This account receivable helps to estimate the upcoming income.

Let‘s take a look the part of balance sheets.

Cash	$120	Accounts Payable	$60
Marketable Securities	$20	Expenses	$40
Account Receivables	$80	Notes Payable	$10
Inventory	$100	Long Term Debt	$20
Total Current Assets	$320	Total Current Liabilities	$130

The account receivable ist $80.

5.1.8 Current Accounts Payable

How much money owed by its creditors (bank, suppliers, . This account receivable helps to estimate the upcoming expenses.

Let‘s take a look the part of balance sheets.

Cash	$120	Accounts Payable	$60
Marketable Securities	$20	Expenses	$40
Account Receivables	$80	**Notes Payable**	$10
Inventory	$100	Long Term Debt	$20
Total Current Assets	$320	Total Current Liabilities	$130

The account payable ist $10.

5.1.9 Inventory Turnover

- How efficiently a company sells and replaces its inventory during a period of time (daily, monthly, or yearly.
- The ability to generate sales and quickly re-stock.

`Inventory Turnover` = Sales / Inventory or Inventory Turnover = Cost of Goods Sold / Average Inventory

Example:

Berlin’s Paper Company sells office paper. During the current year, Berlin reported the cost of goods sold

on its income statement of $2,000,000. Berlin's beginning inventory was $6,000,000 and its ending inventory was $8,000,000. Berlin's turnover is calculated like this:

Inventory Turnover= 2,000,000/ [(6,000,000+8,000,000)/2)= 0,143

It means that Berlin Paper Company only 14,3% of its inventory during the year.

5.1.10 Budget Variance

The difference between the predicted cost or revenue and actual values. Optimistic forecasting or poor decisions caused significant variance (big variance). This KPI used predictive analytics with mathematical and statistical methods.

`Budget Variance`: Predictive Cost - Actual Cost

Not only the cost but also sales, budget, etc.

5.1.11 Sales Growth

The growth of sales over a certain period (daily, weekly, quarterly, or yearly).

`Sales Growth`= Current Period Net Sales - Prior Period Net Sales) / Prior Period Net Sales * 100

Example:

Sales 2018: $10 Mio Sales 2017: $5 Mio

So, the sales growth is (10-5)/5*100= 100%

A Sales hat increased by 100% than last years.

5.1.12 Days Sales Outstanding (DSO)

The average number of days required for clients to pay for a company. If the DSO is lower, the company can focus on ordering additional supplies.

`Days Sales Outstanding`= (Accounts Receivable / Net Credit Sales) * 365

- Accounts Receivable: $50,000
- Net Credit Sales: $400,000

So, the DSO is [50/400] *365= 45.63

It takes 46 days to collect cash from the customers.

5.1.13 Payment Error Rate

Uncompleted payments due to a lack of approval, poor documentation or a missing reference.

5.1.14 Complete List

- Profit Indicators
  - Earnings before Taxes (EBT)
  - Earnings before Interest and Taxes (EBIT)
  - Earnings before Interest, Taxes and Amortization (EBITA)

  - Profit or Loss from Ordinary Business Operations
  - Profit or Loss from Extraordinary Operations
  - Operating income from Ordinary Business

  - Non-operating income from Ordinary Business

  - Result from Discontinued Operations
  - Non-periodic Income
  - Net Operating Profit After Taxes (NOPAT)
- Profitability Indicators
  - EBIT-Turnover-Yield
  - Return On Sales (ROS)
  - Return On Equity (ROE)
  - Return On Assets (ROA)
  - Earnings Per Share (EPS)
  - Return On Investment (ROI)
  - Return On Invested Capital (ROIC)
  - Return On Capital Employed (ROCE)
  - Return On Net Assets (RONA)
  - Risk Adjusted Return On Capital (RAROC)
  - Cost-Income Ratio (CIR)
- Liquidity Indicators
  - Cash Ratio
  - Quick Ratio
  - Current Ratio
  - Working Capital
  - Cash-burn Rate
- Tests of Solvency
  - Debt-to-Equity Ratio
  - Debt-to-Cash Ratio
  - Interest Coverage Ratio
- Cash Flow Measures
  - Cash Flow
  - Gross or Net Cash Flow
  - Free Cash Flow (FCF)
  - Operating Cash Flow (OCF)
  - Earnings before Interest, Taxes, Depreciation and Amortization (EBITDA)
- Cash Flow Ratios
  - Cash Flow Margin
  - Cash Flow Return on Investment (CFROI)
  - Cash Flow Return On Equity (CFROE)
  - EBITDA-Turnover-Yield
  - Income-Tax Burden Ratio
- Financial Structure Indicators
  - Equity-To-Fixed-Assets Ratio (Level I)
  - Equity-To-Fixed-Assets Ratio (Level II)
  - Equity-To-Fixed-Assets Ratio (Level III)
  - Equity Ratio

  - Financial Leverage Index
- Efficiency Ratios
  - Average Collection Period
  - Average Payment Period
  - Cash-to-Cash Cycle
  - Asset Turnover Ratio
  - Asset Coverage Period
- Value Based Management (VBM)
  - Cash Value Added (CVA)
  - Economic Profit (EP)
  - Economic Value Added (EVA)
  - Weighted Average Cost of Capital (WACC)
- Capital Market Tests
  - Market-to-Book Ratio
  - Stock Yield
  - Dividend Yield
  - Price-Earnings Ratio (P/E Ratio)
  - Price-To-Cash Flow Ratio
  - Cash Flow per Share
- Capital Budgeting Tests
  - Payback Period
  - Time Adjusted or Discounted Payback Period

Sources:

- `Operating Cash Flow [40]`
- `Current Ratio Formula [41]`
- `Quick Ratio [42]`
- `Burn Rate [43]`
- `Net Profit Margin [44]`
- `Working Capital [45]`
- `Current Accounts Receivable [46]`
- `Current Accounts Payable [47]`
- `Inventory Turnover Ratio [48]`
- `Budget Variance [49]`
- `Sales Growth [50]`
- `Days Sales Outstanding - DSO [51]`
- `ControllingKennzahlen Key Performance Indicators Zweisprachiges Handbuch Deutsch/Englisch [52]`

5.2 Customer Perspective

According to research from `walkerinfo.com` (Customer Research Consulting), customer satisfaction will overtake as the key success. The company requires volumes of customers, number contacts, and number employees, and other data as a key factor. It depends on the situation and the company business model.

Below is an example of the most using KPI for Customer Perspective:

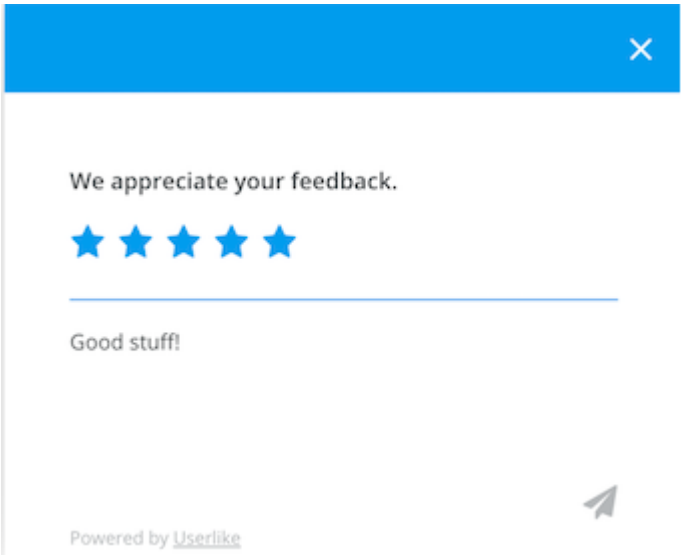


Figure 5.3:

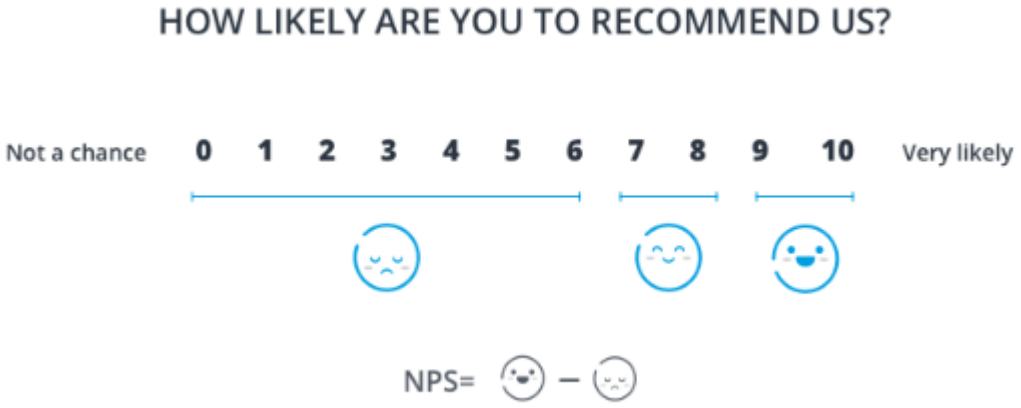


Figure 5.4:

5.2.1 Customer Satisfaction Score (CSAT)

Measuring CSAT is very hard. Customers need to express an emotion, and emotions are harder to indentify. CSAT can consist of numbers or some icon (like stars, smiley faces, color, etc).

Example: Users uses a 5 stars scale for ratings. (See Figure 5.3)

5.2.2 Net Promoter Score (NPS)

How likely customers are referring the products to someone else.

Example: Company ask the customers how likely they are to recommend the products on a scale from 1 to 10. (See Figure 5.4)

5.2.3 First Response Time

Customers are changed like the Spice Girls: “If you wanna get with me, better make it fast!”. The Customers expect an excellent shopping experience. A Salesforce study (Customer Relationship Software Company)

Figure 5.5:

found that a third of the respondents felt positive about companies that offered a quick first response. First Response Time is calculated by subtracting the time of the request from the time of the initial reply. (See Figure 5.5)

5.2.4 Customer Retention Rate

The ability to keep a customer over a set period of time (daily, monthly, yearly).

`Customer Retention Rate` = ((CE – CN) / CS)) x 100

- CE = The customers (end of the period)
- CN = The new customers (acquired during a period)
- CS = The customers (the start of the period)

5.2.5 Employee Engagement

Based on `Harvard Business Review` (https://bit.ly/2bei5oQ), "Long - term employment relationships are key to employee motivation at high performance and sustainability levels." A high employee turnover, on the other hand, will cost you up to twice an employee's salary for finding and training new employees. (See Figure 5.6)

This KPI about employees satisfication. The standars approach ist direct asking. Another method is survey. The survey should consist of the following engagement levels:

- Management quality and time investment
- Influence from colleagues
- Relationships
- Work schedule

Another example of customer perspektive are:

- Annual sales/customers($)
- Average custome size($)
- Customer rating(%)
- Average time from customer contact to sales response(No)
- Average time spent on customer relations(No)
- Customers/employee (No or %)
- Satisfied-customer index(%)
- Customer-loyalty index(%)
- Market Share

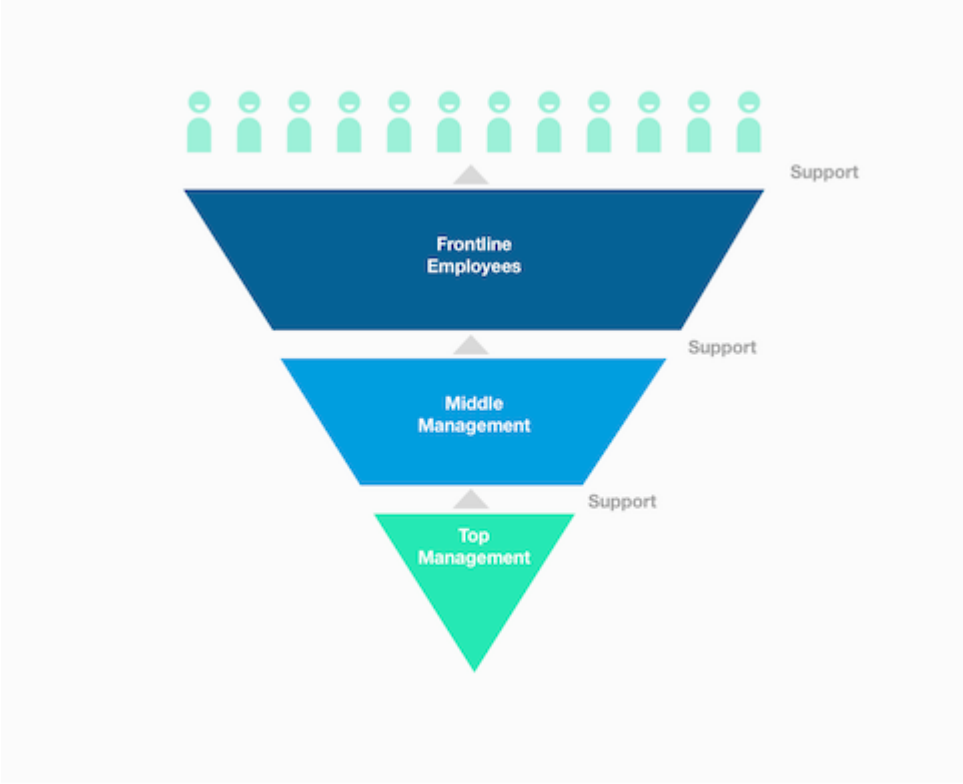

Support
Frontline Employees
Support
Middle Management
Support
Top Management
Successful organizations provide a positive, rewarding work environment.

Figure 5.6:

- No of Customer Complaints
- Return Rates
- Response Time
- Cost/customer($)
- Customers Lost(No or %)
- Customer retention
- Number of customers
- Annual sales per customer
- Marketing cost as a % of sales(%)
- Marketing expenses($)
- Number of proposals made
- Brand-image index (%)
- Response rate
- Sales volume
- Sales per channel
- Average customer size
- Customers per employee
- Frequency of sales transactions
- Sales closed/sales contacts(%)
- Number of visits to customers(No)
- Service expense/customer/year($)

5.2.6 Complete List

- Customer Relationship Management (CRM)
  - Customer Acquisition Rate
  - Customer Churn Rate
  - Customer Retention Period
  - Customer Significance Level
  - Cross-Selling Ratio
  - Customer Lifetime Value (CLV)
  - Customer Satisfaction Index
  - Customer Complaint Ratio
  - Flop Rate
- Marketing Communication Indicators
  - Media Coverage Level
  - Click Through Rate (CTR)
  - Conversion Rate
  - Cost per Thousand
  - Brand Awareness Level
- Product Pricing
  - Profit Margin
  - Gross Margin
  - Absolute Contribution Margin
  - Percentage Contribution Margin

  - Price Reduction Rate
  - Direct Product Profit (DPP)
  - Price Elasticity of Demand (PEoD)
  - Purchasing Power Index
- Cost-Profit-Volume Analysis
  - Break-Even Point (BEP)
  - Margin of Safety
  - Margin of Safety-Factor
  - Cash Point
- Market Coverage Indicators
  - Internationalization Level
  - Distribution Coverage Level
  - Customer Coverage Ratio
  - Market Saturation Level
- Market Position Indicators
  - Absolute Market Share
  - Relative Market Share
  - Bid Acceptance Rate
- Sales Efficiency Indicators
  - Sales per Reference Parameter
  - Contribution Margin per Reference Parameter
  - Sales Space Productivity
  - Capacity Coverage Ratio
  - Book-to-Bill Ratio
  - Finished Goods Turnover Period

Sources:

- `ControllingKennzahlen Key Performance Indicators Zweisprachiges Handbuch Deutsch/Englisch [52]`
- `The 6 Customer Service KPIs You Should Be Tracking [53]`
- `First Response Time [54]`
- `Germany's Midsize Manufacturers Outperform Its Industrial Giants [55]`
- `14 Key Performance Indicators (KPIs) to Measure Customer Service [56]`

5.3 Process Perspective

5.3.1 Cost performance index (CPI)

The financial effectiveness and efficiency of a project. For example, if a project has an earned value of £30,000 but actual costs were £12,000.

`Cost performance index` = EV / AC = 30,000 / 12,000 = 2.5

Interpretation of Results:

- CPI < 1 : over budget.
- CPI $= 1$: on budget.
- CPI > 1 : under budget.

Example:

- CPI = 0 : the project has not started.
- CPI = 0.5 : the project has spent twice the amount that it should have at this point.
- CPI = 1.0 : the project is on schedule.
- CPI = 2.0 : the project has spent half the amount that it should have at this point.

If the ratio has a value higher than 1 then it indicates the project is performing well.

5.3.2 Schedule Performance Index (SPI)

Indicates how efficiently the company to the planned project schedule. It is the efficiency of the time on the project.

`Schedule Performance Index` = (Earned Value) / (Planned Value)

Example:

A Company has a 12 months project. The budget of the project is 200,000 $. 120,000 $ has been spent in six months and the company finds that only 40% of the work has been completed.

Given in the question:
Actual Cost (AC) = $120,000 Planned Value (PV) = 50% of $200,000 = $100,000 Earned Value (EV) = 40% of $200,000 = $80,000 And then, Schedule Performance Index (SPI) = EV / PV = 80,000 / 100,000 = 0.8

The Schedule Performance Index is 0.8. This Company is behind schedule since the Schedule Performance Index is less than one.

5.3.3 Complete List

- Project Controlling
  - Schedule Performance Index (SPI)
  - Cost Performance Index (CPI)
  - Time Estimation at Completion (TEAC)
  - Estimate at Completion (EAC)
  - To-Complete-Performance Index (TCPI)
  - Process Acceleration Costs
- Quality Controlling
  - Quality Rate
  - Rejection Rate

  - Follow-up Costs Ratio
  - Conformity Costs Ratio
  - Non-Conformity Costs Ratio
- Supply Chain Management
  - Procurement Efficiency Ratio
  - Supply Chain Cycle Time
  - Faulty Incoming Delivery Rate
  - Faulty Outgoing Delivery Rate
  - Vertical Integration Level
  - Supplier's Service Level
  - Cooperation Index
- Production Capacity Management
  - Plant Availability Time
  - Plant Downtime Rate 236
  - Maintenance Cost Intensity
  - Capacity Utilization Level
  - Contribution Margin per Unit of the Constrained Resource
- Process Controlling
  - Throughput Time (TPT)
  - Days Inventory Outstanding (DIO)
  - Inventory Turnover Ratio
  - Material Coverage Period
  - Process Cost Rate
  - Expected Process-based Loss
  - Machine Hour Rate
  - Bottleneck-induced Incremental Costs
- Sustainability Management
  - Resource Consumption Level
  - Resource Consumption Efficiency Level
  - Sustainable Value
  - Emission Volume of Production-related Pollutants
  - Emission Volume of Usage-related Pollutants
  - Disposal Costs Ratio
  - Recycling Ratio

Sources:

- `ControllingKennzahlen Key Performance Indicators Zweisprachiges Handbuch Deutsch/Englisch`

[52]
- Cost performance index (CPI) [57]
- Cost Performance Index - Earned Value Management [58]
- Schedule Performance Index (SPI) & Cost Performance Index (CPI) [59]

5.4 Human Resource and Innovation Perspective

Retention

The formula is the following:

Retention :(The number of employees who stayed at the company for the whole time period)/(The number of employees at the start of the time period) X 100

Example: If 100 people were working at my company as of January 1st and 90 of those same people were still working at my company as of January 30th, my retention rate for the month of January would be the following: 90/100 * 100 = 90%

Time in The Position

How long are employees in the same position

Absenteeism

- Delays
- Sick leave
- Excused or unexcused absences

Recruitment Time

The time between employee leaving and another candidate selected to replace him.

Education

The courses for an employee that has a direct impact on the company performances.

Time to Achieve Goals

The efficiency of the workforce to see how long it takes to finish the tasks.

Accidents at Work

The number of accidents in the workplace.

Complete List:

- Personnel Cost Management
  - Personnel Costs Ratio
  - Supplementary Personnel Costs Ratio
  - Personnel Costs per Employee
  - Unit Labor Cost
- Human Resource Controlling
  - Labor Productivity
  - Overtime Quota
  - Workforce Composition Ratios
  - Internally-staffed Executive Positions Ratio
  - Staff Recruitment Period
- Human Resource Development Indicators

  - Apprenticeship Quota
  - Trainee Absorption Rate
  - Professional Development Training Time per Employee
  - Professional Development Training Costs Ratio
- Organizational Behavior Indicators
  - Labor Turnover Rate
  - Employees Satisfaction Index
  - Sickness-Absenteeism Rate
  - Accident Occurrence Rate
  - Participation Rate in Ideas Management
- Innovativeness
  - Innovation Rate
  - Research and Development Intensity (R&D Intensity)
  - Research and Development Costs Ratio (R&D Costs Ratio)
  - Break-Even Time

Sources:

- `ControllingKennzahlen Key Performance Indicators Zweisprachiges Handbuch Deutsch/Englisch [52]`
- `7 KEY INDICATORS OF HUMAN RESOURCES - HR KPI [60]`

5.5 By Industry

Source: `KPI Examples - Performance management starts with figuring out what to measure [62]`

5.5.1 T & L Industry

Example KPIs for the Transportation and Logistic (Warehousing) Industry:

- Annualized inventory turns: a ratio showing how many times a company has sold and replaced inventory during a given period (years, months, 10 years).
- Annualized cost of goods sold (COGS)/average daily inventory value
- Backlog value
- Value of open, not yet fulfilled, booked order lines
- Book to fulfill ratio
- Booked order value/fulfilled value
- Book to ship days
- Average of shipped date - Firm date (booked date used if no firmed date)
- Booked order value
- Booked order line value (not including returns)
- Claims percentage for freight costs
- Customer order promised cycle time
- Defects per million opportunities
- Inventory months of supply
- On-time line count
- On-time pickups
- Pick exceptions rate

- Percentage of picks with exceptions
- Pick release to ship
- Planned inventory turns
- Planned cost of goods sold/planned inventory value
- Planned margin
- Planned revenue - Planned costs
- Planned margin percentage
- Planned margin/planned revenue
- Planned on-time shipment
- Planned service level (percentage of shipments shipped on time)
- Planned resource utilization
- Planned resource usage
- Product revenue
- Product sales revenue (not including service) recognized in selected period (based on AR invoice lines)
- Product revenue backlog
- Value of booked order lines less returns plus deferred revenue backlog (invoiced but not recognized) Production value
- Value of work-in-process (WIP) completions into inventory
- Production to plan rate
- Production standard value/planned standard value
- Receipt to put-away
- Time elapsed from pick release to ship confirm
- Time elapsed from receipt
- Transit time

5.5.2 Wholesale Trade

Example KPIs for the Wholesale Trade Industry:

- Dock turnaround time
- Freight costs (minimize costs without affecting deliveries)
- Inventory accuracy, stockouts
- Inventory carrying costs
- Inventory turns per year
- Logistics costs per year
- Low-velocity inventory comparison through sectors
- Order fill rate and accuracy
- Technology used to execute inventory strategies
- Warehouse flow-through (or some measure of yard or warehouse productivity)
- Wholesale revenue
- Total factor productivity
- Labor productivity
- Return on assets
- Profit margin
- Debt to equity
- Inventory turnover
- Asset utilization
- Collection efficiency

5.5.3 Utilities Industry

Example KPIs for the Utilities Industry:

- Annual labor cost per device
- Average cost per job category

- Average cost per megawatt produced
- Average labor hours per device per year
- Average maintenance cost per mile of pipe/line/cable
- Average number of days each work order is past due
- Average number of labor hours to complete a maintenance task
- Average response time to fix breaks
- Average revenue per megawatt produced
- Average time to settle a rate case
- Consumption analyzed by units consumed and target reduction achieved
- Crew productivity
- Drinking water quality - Percentage of water tests that meet regulatory standards
- Electrical grid load
- Equipment failure rate
- Equipment unavailability, hours per year - Planned maintenance
- Equipment unavailability, hours per year - Sustained fault
- Equipment unavailability, hours per year - Temporary fault
- Equipment unavailability, hours per year - Unplanned maintenance
- Maintenance backlog
- Maintenance cost as a percentage of manufacturing cost
- Maintenance technician's skill level improvement, year over-year
- Mean time to repair
- Number of complaints received by type
- Number of customers who were cut off due to violations of regulations
- Number of disconnections
- Number of pending work orders
- Number of power failures per year
- Number of reported gas leakages per 1,000 households
- Number of sewage blockages per month/year
- Number of staff per 1,000 customer connections
- Number of uncontrolled sewage overflows affecting private properties
- Outage time per event
- Percentage of customers that would characterize their bills as accurate and timely
- Percentage of possible power revenue billed
- Percentage reduction in number of complaints to the local regulatory body
- Percentage reduction in number of employee injuries
- Percentage reduction in number of equipment failures
- Percentage of maintenance work orders requiring rework
- Percentage of man-hours used for proactive work
- Percentage of scheduled man-hours to total man-hours
- Profit redistribution (rural electric coops)
- Reduction in hazardous liquid spill notification time
- Reduction or stabilization in rates (municipally owned utilities)
- Response time to gas or water leaks
- Sewage system reliability
- Station unavailability - Planned maintenance
- Station unavailability - Sustained fault
- Station unavailability - Temporary fault
- Total shareholder returns (investor-owned utilities)
- Total time to complete new customer connections
- Transformer/pump station reliability
- Voltage deviations per year
- Water system reliability

5.5.4 Retail Industry

Example KPIs for the Retail Trade Industry:

PRODUCT SALES

- Average inventory
- Cost of goods sold
- Gross profit budget percentage
- Sales budget percentage
- Discount
- Gross profit
- Gross profit and prognostics
- Gross profit and prognostics percentage
- Gross profit budget
- Gross profit campaign
- Gross profit percentage KPI
- Gross profit prognostics
- Gross profit prognostics percentage
- Gross profit standard
- Gross profit year to date
- Number of stores
- Product quantity
- Sales
- Sales and prognostics
- Sales campaign
- Sales growth period
- Sales growth year
- Sales growth year by week
- Sales prognostics
- Sales standard
- Sales trend percentage KPI
- Sales value-added tax (VAT)
- Sales view
- Sales view year-to-date
- Share prognostics
- Time range

FINANCE AND ACCOUNTING

- Accounts payable turnover
- Accounts receivable turnover days
- Acid test ratio
- Administrative cost percentage
- Break-even (dollars)
- Cash conversion cycle
- Contribution margin
- Cost of goods
- Cost of goods sold
- Current ratio
- Ending inventory at retail
- Gross margin
- Gross margin return on investment
- Initial mark-up
- Interest cost percentage
- Inventory turnover

- Maintained mark-up (dollars)
- Margin percentage
- Mark-up percentage
- Net receipts
- Net sales
- Retail price
- Return on capital invested
- Sales per square foot
- Stock turnover days
- Total asset sales ratio
- Turnover

SALARY

- Real absence hours
- Real absence share
- Real GPWH
- Real overtime hours
- Real overtime share
- Real TWH
- Real working hours
- Salary
- Salary amount
- Salary amount exchange currency
- Salary hours
- Salary turnover share

SALARY TARGETS

- Real absence hours
- Real GP work hours
- Real total work hours
- Salary absence percentage
- Salary GP work hour
- Salary overtime percentage
- Salary target absence percentage
- Salary target GP work hour
- Salary target overtime percentage
- Salary target turnover percentage
- Salary target work hour
- Salary turnover percentage

HOURLY SALES

- Customers per hour
- Discount
- Gross profit
- Items
- Margin per customer
- Number of customers
- Sales growth year
- Sales growth year percentage
- Sales last year
- Sales per customer
- Sales trend percentage
- Sales view
- Total number of stores

BUDGET SALES

- Budget gross profit
- Budget number of customers
- Budget sales
- Customers
- Discount
- Gross profit
- Items
- Sales
- Sales exchange currency
- Sales VAT

PAYMENT WITH POINT-OF-SALE (POS) STATISTICS

- Amount
- Amount exchange currency
- Items
- Number of customers
- Number of items
- Refund amount
- Refund count
- Sales income VAT
- Time range
- Transaction cancel amount
- Transaction cancel count
- Transaction cancel percentage
- Void amount
- Void count
- Void percentage
- Zero sale count

HOURLY PRODUCT SALES

- Gross profit percentage
- Item discount
- Item gross profit
- Item quantity
- Item sales
- Item sales exchange currency
- Item sales VAT
- Items sold

5.5.5 Real Estate and Rental and Leasing

Example KPIs for the Real Estate and Rental and Leasing Industry:

REALTOR WEBSITE

- Conversation rate (i.e., take rate) - Number of conversations over number of website visits
- Top conversion page exit - The page where website visitors change their minds and exit your website.
- Traffic source percentage - Website visits referred by

REAL ESTATE OFFICE

- Advertising and promotion
- Average commission per sale
- Average commission per salesperson
- Commission margin

- Net profit
- Office cost (telephone, fax, and other office cost)
- Rent cost of premises
- Sold homes per available inventory ratio
- Total income
- Wages and salaries (including commissions and vehicle allowances)
- Year-to-year variance on average sold price
- Year-to-year variance on dollar volume of sold listings
- Year-to-year variance on sold average dollar per square foot

COMMERCIAL PROPERTY MANAGEMENT

- Annual return on investment in percentage
- Construction/purchaser rate - New constructed or purchased units over time
- Cost per square foot
- Equity value growth in percentage
- Lease events coverage ratio - Number of lease inquiries over number of available units
- Management efficiency - Number of leased spaces over number of staff
- Market share growth
- Monthly return on investment as percentage
- Occupancy cost - Cost per occupied unit
- Operation cost to rent income ratio
- Percentage of rent collected
- Price to income as percentage
- Profitability per square foot
- Real estate demand growth - Market rental demands
- Rented space usage quality - Average number of tenant visits over rented space
- Renting cost - Renting cost per square foot
- Renting return on investment - Rent income over cost
- Revenue per square foot
- Risk metrics as percentage
- Total property management income per property manager
- Usage efficiency - Available renting square feet over number of staff
- Utilization (vacancy) rate - Rented square feet over total square feet, or rented units over total units

REAL ESTATE INVESTOR

- Average gross multiplier for portfolio
- Cost per square foot to value per square foot ratio
- Equity to value ratio
- Gross multiplier per commercial property
- LTV (loan to value) ratio per property
- Mortgage rate index
- Overall LTV (loan to value) ratio for portfolio
- Price per square foot to value per square foot ratio
- Profitability per square foot
- Property value growth (market trend)
- Purchase price-to-appraisal value ratio
- Rental value growth rate ROI (return on investment)

5.5.6 Manufacturing Industry

Example KPIs for the Manufacturing Industry:

- Asset utilization
- Availability
- Avoided cost

- Capacity utilization
- Comparative analytics for products, plants, divisions, companies
- Compliance rates (for government regulations, etc.)
- Customer complaints
- Customer satisfaction
- Cycle time
- Demand forecasting
- Faults detected prior to failure
- First aid visits
- First time through
- Forecasts of production quantities, etc.
- Increase/decrease in plant downtime
- Industry benchmark performance
- Integration capabilities
- Interaction level Inventory
- Job, product costing
- Labor as a percentage of cost
- Labor usage, costs-direct and indirect
- Machine modules reuse
- Maintenance cost per unit
- Manufacturing cost per unit
- Material costing, usage
- Mean time between failure (MTBF)
- Mean time to repair
- Number of production assignments completed in time
- On-time orders
- On-time shipping
- Open orders
- Overall equipment effectiveness
- Overall production efficiency of a department, plant, or division
- Overtime as a percentage of total hours
- Percentage decrease in inventory carrying costs
- Percentage decrease in production-to-market lead-time
- Percentage decrease in scrap and rework costs
- Percentage decrease in standard production hours
- Percentage increase in productivity
- Percentage increase in revenues
- Percentage material cost reduction
- Percentage reduction in defect rates
- Percentage reduction in downtime
- Percentage reduction in inventory levels
- Percentage reduction in manufacturing lead times
- Percentage savings in costs
- Percentage savings in inventory costs
- Percentage savings in labor costs
- Percentage savings in transportation costs
- Planned work to total work ratio
- Predictive maintenance monitoring (maintenance events per cycle)
- Process capability
- Productivity
- Quality improvement (first-pass yield)
- Quality tracking-six sigma
- Reduced time to productivity
- Reduction in penalties

- Savings in inventory carrying costs
- Scheduled production
- Spend analytics
- Storehouse stock effectiveness
- Supplier trending
- Time from order to shipment
- Time on floor to be packed
- Unplanned capacity expenditure
- Unused capacity expenditures
- Utilization
- Waste ration reduction
- Work-in-process (WIP)

INSURANCE

- Average insurance policy size
- Claims
- Combined cost and claims ratio
- Combined ratio
- Current premium versus loss
- Earned premium
- Expense ratio
- Expenses
- Exposure
- Loss adjustment expenses (LAE)
- Loss ratio
- Number of days open of insurance claims
- Number of new insurance policies
- Previous premium versus loss
- Underwriting speed of insurances
- Written premium

5.5.7 Finance and Insurance

Example KPIs for the Finance and Insurance Industry:

FINANCE

- Accounting costs
- Accounts payable
- Accounts payable turnover
- Asset turnover rate
- Average sum deposited in new deposit accounts
- Average value of past due loans
- Cash conversion cycle (CCC)
- Cash dividends paid
- Cash flow return on investments (CFROI)
- Common stock equity
- Cost of goods sold (COGS)
- Cost per hour per lawyer (in-house)
- Creditor days
- Cumulative annual growth rate (CAGR)
- Cycle time to perform periodic close
- Cycle time to resolve an invoice error
- Days payable
- Debt-to-asset ratio

- Debtor days
- Direct costs
- Earnings per share (EPS)
- EBIT
- EBITDA
- Economic value added
- Enterprise value/takeover value
- Fixed costs
- Gross margin on managed assets
- Gross profit
- Gross profit margin
- Indirect costs
- Interest expense
- Interest on net worth
- Invoicing processing costs
- Labor and management cost
- Labor and management earnings
- Legal staff per size of revenue
- Long-term debt
- Marginal costs
- Market share
- Net change in cash
- Net interest margin
- Net new money
- Net profit
- Net profit margin
- Number of budget deviations
- Number of invoices outstanding
- Number of past due loans
- Operating income
- Operating leverage
- Operating margin
- Operating profit margin
- Other current liabilities
- Other noncurrent liabilities
- Percentage of accuracy of periodic financial reports
- Percentage of effectiveness in payables management
- Percentage of budget deviation relative to total budget
- Percentage of electronic invoices
- Percentage of financial reports issued on time
- Percentage of invoices requiring special payment
- Percentage of invoices under query
- Percentage of legal budget spent outside
- Percentage of low-value invoices
- Percentage of payable invoices without purchase order
- Preferred stock equity
- Product turnover ratio
- Profit
- Profit loss due to theft
- Profit margin
- Profit per product
- Quick ratio
- Rate of return on assets
- Rate of return on equity

- Return on assets
- Return on capital employed (ROCE)
- Return on investment (ROI)
- Return to equity
- Revenue
- Revenue per employee
- Sales per share
- Same store sales
- Selling general and administrative (SG&A) expenses
- Share price
- Shares outstanding
- Sharpe ratio
- Short-term debt
- Sortino ratio
- Systems cost of payroll process as a percentage of total payroll cost
- Tier 1 capital
- Total assets
- Total current liabilities
- Total equity
- Total legal spending as a percentage of revenue
- Total liabilities
- Total of uninvested funds
- Total quantity of new deposit accounts
- Total sum deposited in new deposit accounts
- Total value of past due loans
- Variable costs

5.5.8 Construction

Example KPIs for the Construction Industry:

- Number of accidents
- Number of accidents per supplier
- Actual working days versus available working days
- Cash balance - Actual versus baseline
- Change orders - Clients
- Change orders - Project manager
- Client satisfaction - Client-specified criteria
- Client satisfaction product - Standard criteria
- Client satisfaction service - Standard criteria
- Cost for construction
- Cost predictability - Construction
- Cost predictability - Construction (client change orders)
- Cost predictability; Construction (project leader change orders)
- Cost predictability - Design
- Cost predictability - Design and construction cost to rectify defects
- Customer satisfaction level
- Day to day project completion ratio - Actual versus baseline
- Fatalities
- Interest cover (company)
- Labor cost - Actual versus baseline
- Labor cost over project timeline
- Liability ratio (over asset) on current versus completion comparison
- Number of defects
- Outstanding money (project)

- Percentage of equipment downtime
- Percentage of labor downtime
- Percentage of backlogs over project timeline
- Percentage of unapproved change orders
- Productivity (company)
- Profit margin - Actual versus baseline profit margin over project timeline
- Profit predictability (project)
- Profitability (company)
- Quality issues at available for use
- Quality issues at end of defect rectification period
- Ratio of value added (company)
- Repeat business (company)
- Reportable accidents (including fatalities)
- Reportable accidents (non-fatal)
- Return on capital employed (company)
- Return on investment (client)
- Return on value added (company)
- Time for construction
- Time predictability - Construction
- Time predictability - Construction (client change orders)
- Time predictability - Construction (project leader change orders)
- Time predictability - Design
- Time predictability - Design and construction
- Time taken to reach final account (project)
- Time to rectify defects

5.5.9 IT Industry

Example KPIs for the Information Technology Industry:

- Annual cost per reading
- Average cost per article
- Average cost per subscription
- Average dollars per email sent or delivered
- Average order size
- Average quarter-hour audience
- Average revenue per subscription
- Average time spent listening per user (day/week/month/year)
- Bounce rate
- Click to open rate (number of unique clicks/ number of unique opens)
- Click-through rate
- Click-through rate (CTR)
- Conversion rate
- Conversion rate (number of actions/unique click-throughs)
- Conversion rates
- Cost per broadcast hour
- Cost per consumed (by viewers/listeners) hour
- Cost per customer
- Cost per lead, prospect, or referral
- Cost per production hour
- Cost per viewer/listener
- Cost per visitor
- Cost per action (CPA)
- Cumulative audience sessions
- Delivery rate (emails sent, bounces)

- Gross ratings points
- Life cycle cost per reading
- Local content as a percentage of all content
- Net subscribers (number of subscribers plus new subscribers) -(bounces + unsubscribes)
- Number of broadcast hours per day/week/month/year
- Number of or percentage of spam complaints
- Number of orders, transactions, downloads, or actions
- Open rate
- Output per employee (unique first run broadcast hours by employee for each medium)
- Pay per click (PPC)
- Pay per lead (PPL)
- Pay per sale (PPS)
- Percentage of broadcast hours by genre (news/sports/entertainment, etc.)
- Percentage of overhead (non-direct operating costs) against total expenditure
- Percentage of orders, transactions, downloads, or actions of emails sent or delivered
- Percentage unique clicks on a specific recurring Iink(s)
- Referral rate ("send-to-a-friend")
- Site stickiness (number of pages visited per visit)
- Subscriber retention (number of subscribers, bounces, unsubscribes/number of subscribers)
- Total cost per subscription
- Total listener hours (day/week/month/year)
- Total revenue
- Total revenue per subscription
- Unique visitors (total number of unique visitors per day/week/month)
- Unsubscribe rate
- Utilization of production resources
- Value per visitor
- Viewers/listeners for each medium as a percentage of total population
- Website actions (number of visits to a specific web page or pages)
- Website traffic (total page impressions per day/week/month)

5.6 By Business Goal

Source: `INTERACTIVE DASHBOARD EXAMPLES [61]`

When it comes to improving on specific business goals, this KPI got you covered.

5.6.1 Improve response time

Goals: Improve representative of Call Centers. (See Figure 5.7)

5.6.2 Increase profit margins

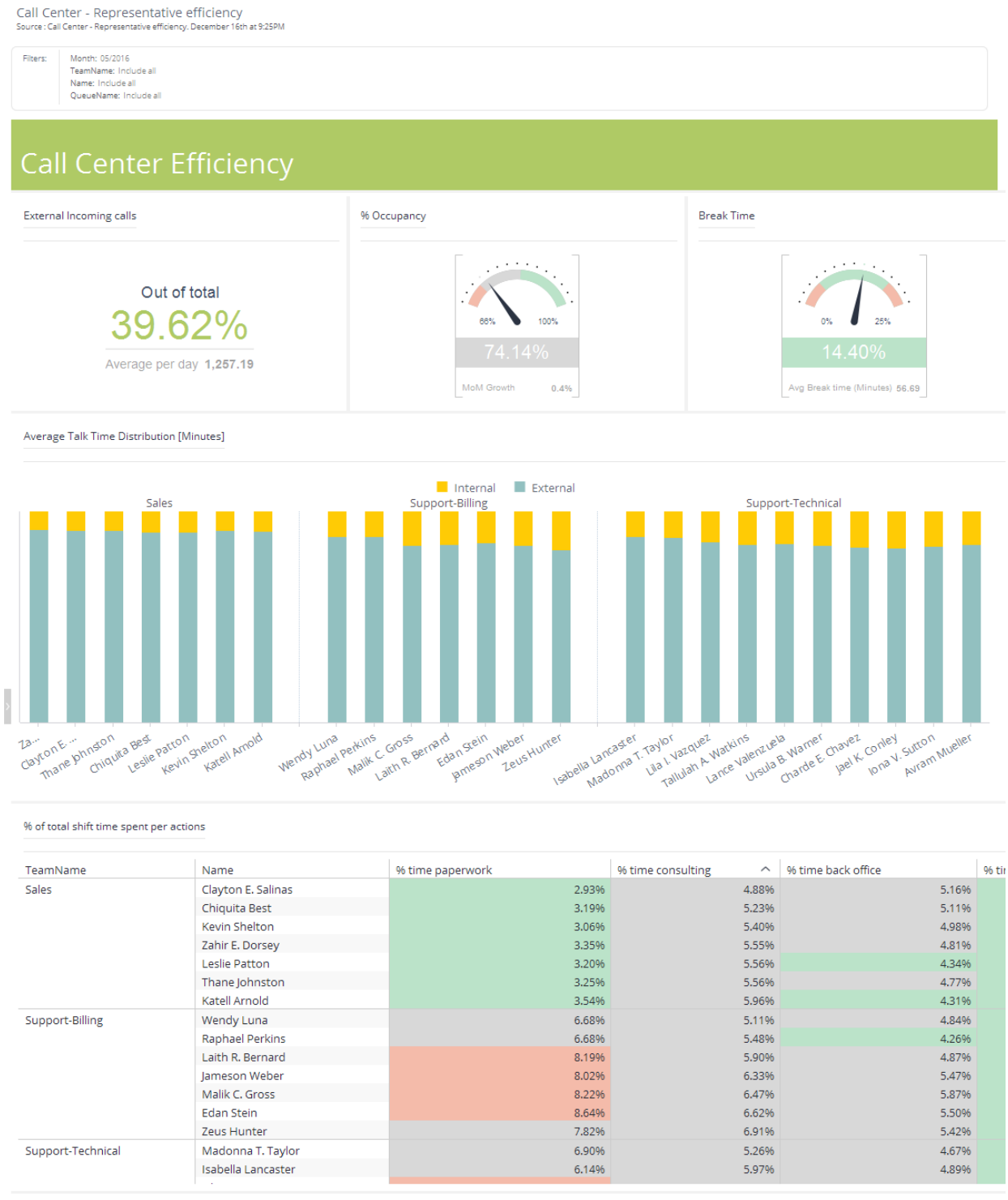
Goals:

- To compare the net profit margin over time
- To compare the net profit margin across industries or subsidiary

(See Figure 5.8)

5.6.3 Optimize campaigns

Goals: Increase sales from online marketing channels. (See Figure 5.9)



TeamName	Name	% time paperwork	% time consulting	% time back office	% ti
Sales	Clayton E. Salinas	2.93%	4.88%	5.16%	
	Chiquita Best	3.19%	5.23%	5.11%	
	Kevin Shelton	3.06%	5.40%	4.98%	
	Zahir E. Dorsey	3.35%	5.55%	4.81%	
	Leslie Patton	3.20%	5.56%	4.34%	
	Thane Johnston	3.25%	5.56%	4.77%	
	Katell Arnold	3.54%	5.96%	4.31%	
Support-Billing	Wendy Luna	6.68%	5.11%	4.84%	
	Raphael Perkins	6.68%	5.48%	4.26%	
	Laith R. Bernard	8.19%	5.90%	4.87%	
	Jameson Weber	8.02%	6.33%	5.47%	
	Malik C. Gross	8.22%	6.47%	5.87%	
	Edan Stein	8.64%	6.62%	5.50%	
	Zeus Hunter	7.82%	6.91%	5.42%	
Support-Technical	Madonna T. Taylor	6.90%	5.26%	4.67%	
	Isabella Lancaster	6.14%	5.97%	4.89%	

Figure 5.7:

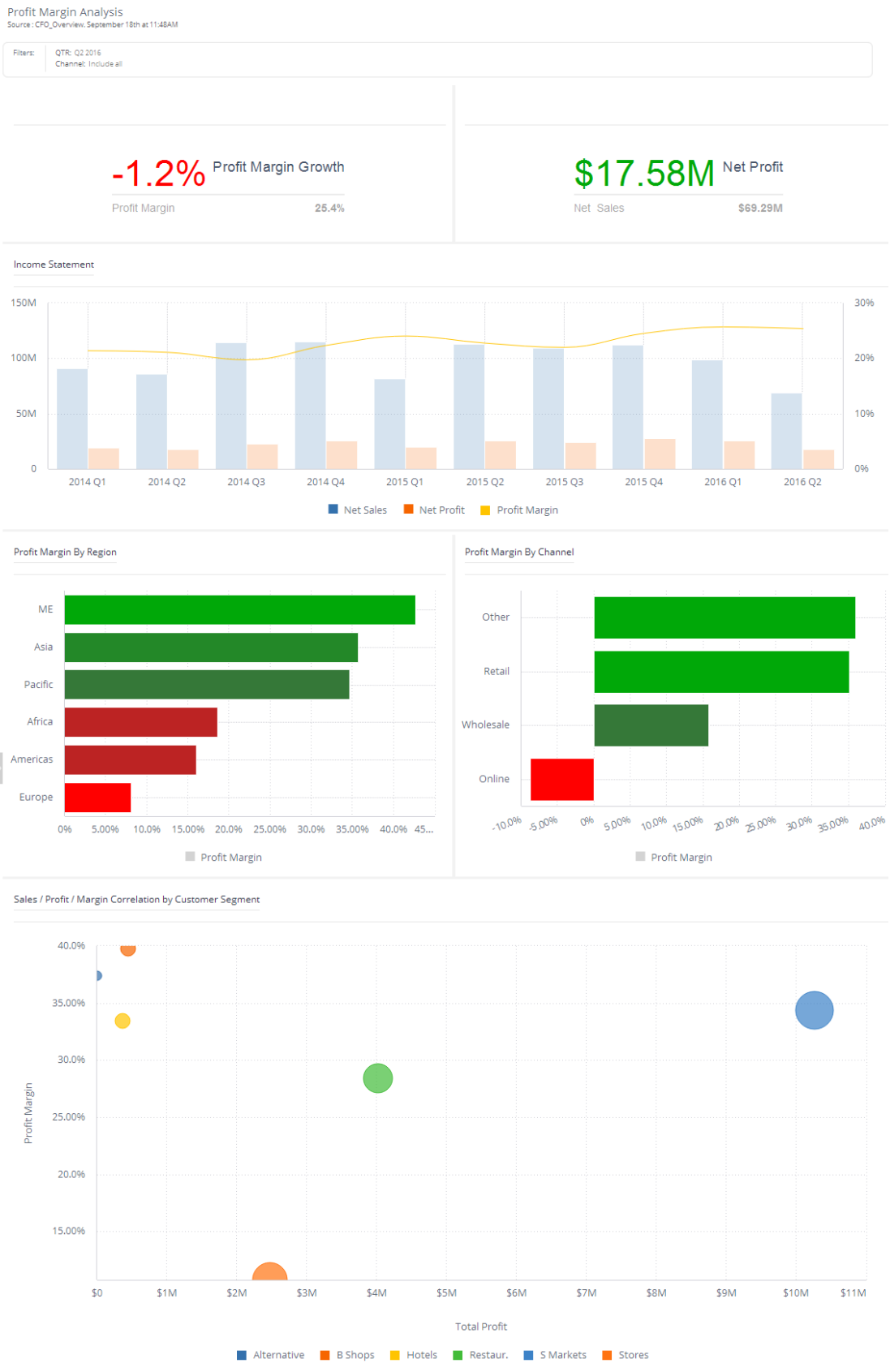
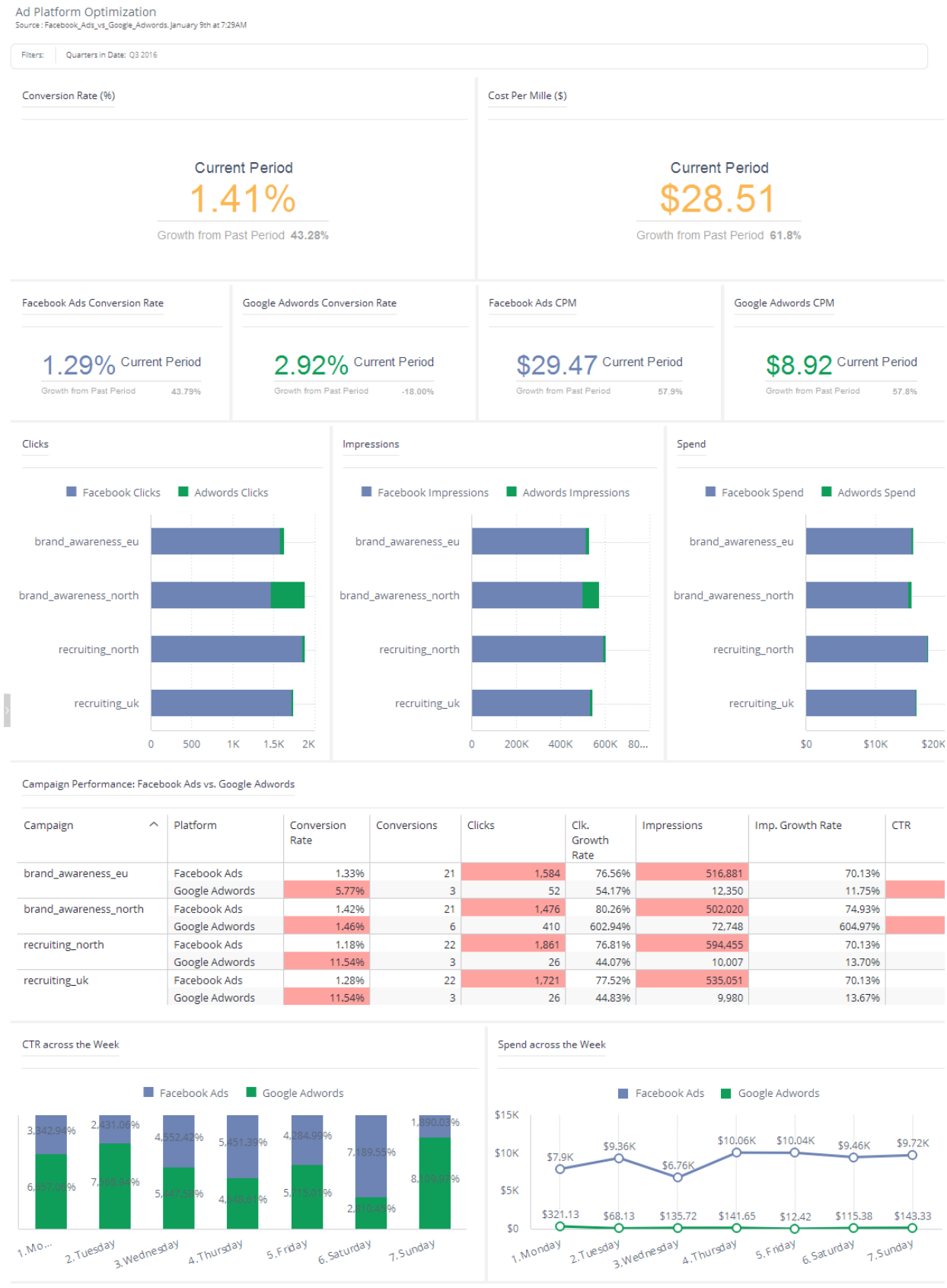

Profit Margin Analysis
Source : CFO_Overview, September 18th at 11:48AM
Filters:
QTR: Q2 2016
Channel: Include all
-1.2% Profit Margin Growth
Profit Margin 25.4%
$17.58M Net Profit
Net Sales $69.29M
Income Statement
150M
100M
50M
0
30%
20%
10%
0%
2014 Q1
2014 Q2
2014 Q3
2014 Q4
2015 Q1
2015 Q2
2015 Q3
2015 Q4
2016 Q1
2016 Q2
Net Sales
Net Profit
Profit Margin
Profit Margin By Region
ME
Asia
Pacific
Africa
Americas
Europe
0%
5.00%
10.0%
15.00%
20.0%
25.00%
30.0%
35.00%
40.0%
Profit Margin
Profit Margin By Channel
Other
Retail
Wholesale
Online
-10.0%
-5.00%
0%
5.00%
10.0%
15.00%
20.0%
25.00%
30.0%
35.00%
40.0%
Profit Margin
Sales / Profit / Margin Correlation by Customer Segment
40.0%
35.00%
30.0%
25.00%
20.0%
15.00%
Profit Margin
$0
$1M
$2M
$3M
$4M
$5M
$6M
$7M
$8M
$9M
$10M
$11M
Total Profit
Alternative
B Shops
Hotels
Restaur.
S Markets
Stores

Figure 5.8:



Campaign	Platform	Conversion Rate	Conversions	Clicks	Clk. Growth Rate	Impressions	Imp. Growth Rate	CTR
brand_awareness_eu	Facebook Ads	1.33%	21	1,584	76.56%	516,881	70.13%	
	Google Adwords	5.77%	3	52	54.17%	12,350	11.75%	
brand_awareness_north	Facebook Ads	1.42%	21	1,476	80.26%	502,020	74.93%	
	Google Adwords	1.46%	6	410	602.94%	72,748	604.97%	
recruiting_north	Facebook Ads	1.18%	22	1,861	76.81%	594,455	70.13%	
	Google Adwords	11.54%	3	26	44.07%	10,007	13.70%	
recruiting_uk	Facebook Ads	1.28%	22	1,721	77.52%	535,051	70.13%	
	Google Adwords	11.54%	3	26	44.83%	9,980	13.67%	

Figure 5.9:

5.7 By Department

Source:KPI Examples - Performance management starts with figuring out what to measure [62]

5.7.1 Sales

Example KPIs for Sales Departments:

- Actual calls
- Actual sales value versus initial bid
- Age of sales forecast
- Average administrative time per sales person
- Average deal size
- Average number of activities (calls, meetings, etc.) to close a deal
- Average price discount per product
- Average price discount per sales person
- Average revenue per product
- Call quota
- Closed sales
- Closing ratio
- Customer acquisitions costs as a percentage of sales value
- Customer churn ratio
- Customer loyalty
- Customer purchase frequency
- Customer satisfaction
- Frequency of sales transactions
- Gross margin per product
- Gross margin per sales person
- New sales person ramp-up time
- Number of certified partners
- Number of deals per partner
- Number of sales orders by FTE
- Number of sales people meeting their quota
- Number of units sold per day/week/month/quarter/year
- Partner churn ratio
- Partner profit margin
- Percentage of converted opportunities
- Percentage of online sales revenue
- Percentage of sales due to launched product/services
- Percentage of sales representatives to achieve quota
- Percentage of sales revenue via partner channel
- Pipeline by sales stage
- Qualified leads
- Qualified opportunities
- Revenue per sales person
- Sales capacity
- Sales cycle time
- Sales per department
- Sales person turnover
- Sales quota
- Time utilization
- Unweighted sum of deal size in sales pipeline
- Value of sales lost
- Win/loss ratio percentage

5.7.2 Marketing

Example KPIs for Marketing Departments:

- Ad click-through ratio (CTR)
- Average response rates of campaigns
- Brand awareness percentage
- Brand consideration
- Brand credibility
- Brand strength
- Column inches of media coverage
- Consumer awareness
- Contact rate (number of contacts effectively contacted / number of contacts in the target list)
- Cost per converted lead
- Cost per lead
- Cost per mille (CPM)
- Delivery of materials
- Effective reach
- Gross rating point (GRP)
- Growth sustainability rate of brand
- Leads generated
- Marketing budget awareness-demand ratio
- Marketing budget ratio (MER)
- Number of article placements in trade magazines
- Number of client visits
- Number of product focus groups conducted
- Number of customer satisfaction surveys administered
- Number of placements in trade magazines
- Number of trade shows attended / participated in
- Percentage of customers willing to promote your product/service
- Q score (a way to measure the familiarity and appeal of a brand, etc.)
- Response rate
- Return on investment (ROI) of brand
- Return on marketing investment (ROMI)
- Revenue generation capabilities of brand
- Staying in budget
- Target rating point
- Total cost of customer acquisition
- Transaction value of brand
- Website click-throughs
- Website hits
- Website leads generated

5.7.3 Finance

Example KPIs for Finance Departments:

- Accounting costs
- Accounts payable turnover
- Accounts receivable collection period
- Accounts receivable turnover
- Actual expenses
- Amount due (per customer)
- Average customer receivable
- Average monetary value of invoices outstanding
- Average monetary value of overdue invoices

- Average number of trackbacks per post
- Budget variance for each key metric
- Budgeted expenses
- Capital expenditures
- Cash conversion cycle (CCC)
- Cash flow return on investments (CFROI)
- Cost of goods sold (COGS)
- Cash dividends paid
- Cost per pay slip issued
- Creditor days
- Current receivables
- Cumulative annual growth rate (CAGR)
- Cycle time for expense reimbursements
- Cycle time to process payroll
- Cycle time to resolve an invoice error
- Cycle time to resolve payroll errors
- Days payable
- Debtor days
- Direct cost
- Discounted cash flow
- Earnings before interest and taxes (EBIT)
- Earnings before interest, taxes, depreciation (EBITDA)
- Economic value added (EVA)
- Employee available time
- Employee scheduled time
- Employee work center loading
- Enterprise value/ takeover value
- Expense account credit transactions
- Expense account debit transactions
- Expense account transactions
- Fixed costs
- Gross profit
- Gross profit margin
- Indirect costs
- Inventory turnover
- Inventory value
- Invoice processing costs
- Internal rate of return (IRR)
- Market share gain comparison percentage
- Net change in cash
- Net income
- Net present value (NPV)
- Number of invoices outstanding
- Number of unapplied receipts
- Number of past-due loans
- Open receivables
- Open receivables amount (per customer)
- Operating leverage
- Past-due receivables
- Payables turnover
- Payment errors as a percentage of total payroll disbursement
- Percentage accuracy of financial reports
- Percentage of bad debts against invoiced revenue
- Percentage of electronic invoices

- Percentage in dispute (per customer)
- Percentage of invoices being queried
- Percentage of invoices requiring special payment
- Percentage of low-value invoices
- Percentage of open receivables (per customer)
- Percentage of payable invoices without purchase order
- Percentage of service requests posted via web (self-help)
- Perfect order measure
- Quick ratio
- Receivables
- Receivables turnover
- Return on capital employed (ROCE)
- Sales growth
- Share price
- Systems cost of payroll process as a percentage of total payroll cost
- Total payables
- Total energy used per unit of production
- Total receivables
- Total sales
- Unapplied receipts
- Variable costs
- Weighted days delinquent sales outstanding
- Weighted days delinquent sales outstanding (per customer)
- Weighted terms outstanding
- Weighted terms outstanding (per customer)

5.7.4 Human Resources

Example KPIs for Human Resources (HR) Departments:

- Actual versus budgeted cost of hire
- Annualized voluntary employee turnover rate
- Annualized voluntary turnover rate
- Average headcount of employees each human resources (HR) employee working is caring for
- Average interviewing costs
- Average length of placement in months for the manager
- Average length of service of all current employees
- Average length of service of all employees who have separated
- Average months placement
- Average number of training hours per employee
- Average number of vacation days per employee
- Average performance scores of departing employees
- Average retirement age
- Average salary
- Average salary for all employees reporting to the selected manager
- Average sourcing cost per hire
- Average time employees are in same job/ function
- Average time to competence
- Average time to update employee records
- Average training costs per employee
- Compensation cost as a percentage of revenue
- Contingent workers
- Employee satisfaction with training
- End placements
- Female to male ratio

- Full-time employees (FTEs) per human resources (HR) department FTE
- Headcount of contingent workers for the manager
- HR average years of service (incumbents)
- HR average years of service (terminations)
- HR department cost per FTE
- HR headcount - Actual
- HR headcount - Available
- HR to employee staff ratio
- Job vacancies as a percentage of all positions
- New hire quality
- Time to fill
- Hiring manager satisfaction
- Cost per hire
- Staffing efficiency
- Internal, external, and total headcount recruiting costs and ratios
- Number of end placements made in the reporting period for the manager
- Part-time employees as a percentage of total employees
- Percentage of employees receiving regular performance reviews
- Percentage of employees that are near or at max for their vacation balances
- Percentage of HR budget spent on training
- Percentage of new hire retention
- Ratio of internal versus external training
- Ratio of standard level wage to local minimum wage
- Return on investment (ROI) of training
- Total overtime hours as a percentage of all work hours
- Training penetration rate (percentage of employees completing a course compared to all FTEs)
- Workforce stability

5.7.5 Information Technology

Example KPIs for Information Technology (IT) Departments:

- Account create success
- Account termination success
- Active directory performance index
- Alert-to-ticket ratio
- Average data center availability
- Call center PBX availability
- Campus PBX availability
- Customer connection effectiveness
- Data center capacity consumed
- Email client availability
- Exchange server availability
- Incidents from change
- Internet proxy performance
- Network availability - High availability sites
- Network availability - Standard sites
- Network manageability index
- No problem found/duplicate tickets
- Percentage of branch office backup success
- Percentage of circuits exceeding target utilization
- Percentage of IT managed servers patched at deadline
- Percentage of production servers meeting software configuration standards
- Percentage of security update restarts within maintenance window
- Percentage successful remote access server (RAS) connections

- Phone answer service level
- Priority 1 and priority 2 network incidents meeting SLA
- Product adoption status and compliance
- Restore success rate
- Server growth rate
- Server manageability index
- Service desk client satisfaction - Percentage dissatisfied
- Service desk tier 1 resolution rate
- Service desk time to escalate
- Service desk time to resolve
- Storage utility service availability
- Storage utility utilization
- Virtual machine provisioning interval
- Virtual server utility availability
- Web server availability

5.7.6 Customer Service

Example KPIs for Customer Service Departments

- Agent's full-time employees (FTEs) as percentage of total call center FTEs
- Answering percentage (number of sales calls answered/total number of sales calls offered)
- Average after-call work time
- Average number of calls/ service request per handler
- Average queue time of incoming phone calls
- Cost per minute of handle time
- Costs of operating call center/ service desk
- Email backlog
- Field service technician utilization
- Hit rate (products sold compared to total received sales calls)
- Inbound abandon rate
- Inbound agent dialed calls
- Inbound availability rate
- Inbound average talk time
- Inbound average wrap time
- Inbound call center leads created
- Inbound call center opportunities created
- Inbound calls handled
- Inbound calls handled per agent hour
- Inbound service level
- Number of complaints
- Percentage of customer service requests answered in given timeframe
- Percentage of calls transferred
- Total calling time per day/week/month

Chapter 6

Deutsche Bahn Use Case

DB Headquarters has the consulting function for all of DB subdiaries. DB Headquarters works together with the all data sources to devise data usage strategies. By collaborating with the domain's experts, DB Headquarters develops matching mathematical models and all relevant KPI (Key Performance Indicators). The data and corresponding analytical results are then prepared for visual presentation and made available for further examination. The DB company created KPI with Microsoft Power BI (one of the KPI solution).

At Deutsche Bahn the using of KPI have three functions:

1. General Overview
2. Warning system
3. Decisions making

Below are some of use the case in Deutsche Bahn Headquarters:

6.1 Spend Analysis KPI: IT Department

6.1.1 Overview

The IT Spend analysis is the process of searching the company aggregate spend throughout the IT systems. During a analysis, the company identify, gather, categorise, and analyse everything a company spends, and to find areas where the company can save money and increase efficiency. It analyze the planned vs. actual costs of an IT systems. This KPI Dashboard helps the company understand how well the company planned for the year and investigate areas with huge deviations from the plan.

6.1.2 Dataset

This is real data from obviEnce (http://obvience.com/) that has been anonymized. Find the data in my github repository: https://github.com/itsmecevi/spend-analysis-it

This analysis contains 8 entity (table) and every table has an attribute (parameters as a columns).

1. business area

- Attribute: Business Area, Business Area ID

2. cost element

- Attribute: Cost element name, Cost Element Group, Cost Element Sub Group, Cost Element ID

3. country region

- Attribute: Sales Region, Country/Region, Country/Region ID

4. date

- Attribute: Date, Year, Period, Month

5. department

- Attribute: VP, Department

6. it-area

- Attribute: IT Area, IT Sub Area, IT Sub Area ID

7. scenario

- Attribute: Scenario, Scenario ID, ScenarioDescription

8. fact_it

- Attribute: Date, Value, Department, Cost Element ID, Country/Region ID, Business Area ID, IT Sub Area ID, Scenario ID

6.1.3 Glossary

YTD: Year-to-date ,is a period, starting from the beginning of the current year.

Var: Variance or the deviation

LE: Latest Estimate

6.1.4 Solution

IT Spend Trend Dashboard: YTD IT Spend Trend Analysis page:

- Var Plan % by Sales Region
- Var Plan by Month
- Var Plan , Var Plan % and Actual by Business Area and Period
- IT Area

Spend by Cost Element Dashboard: YTD Spend by Cost Elements page:

- Plan and Target
- Var Plan % and Var LE3 % by IT Area
- Var Plan % by Sales Region
- Amount by Month and Scenario

Plan Variance Analysis Dashboard:

- Variance Latest Estimates
- Var Plan % by Business Area
- Var Plan % by Business Area:
- Var Plan by Sales Region and country
- Var Plan % by Month and Business Area

6.1.5 Technical documentation

Below are the steps by step to make a KPI Dashboard for this solution.

1. Exploratory Data Analysis with R Programming:

See the section `4.2 EDA: Exploratory Data Analysis` for doing the data cleaning. The dataset based on `Star Schema`. We need just identify the `fact` table. (See the section `3.1 : Stage Four - Data Modeling` for more info about `fact` table). Below are the `Exploratory Data Analysis` results.

with `dlookr package`:

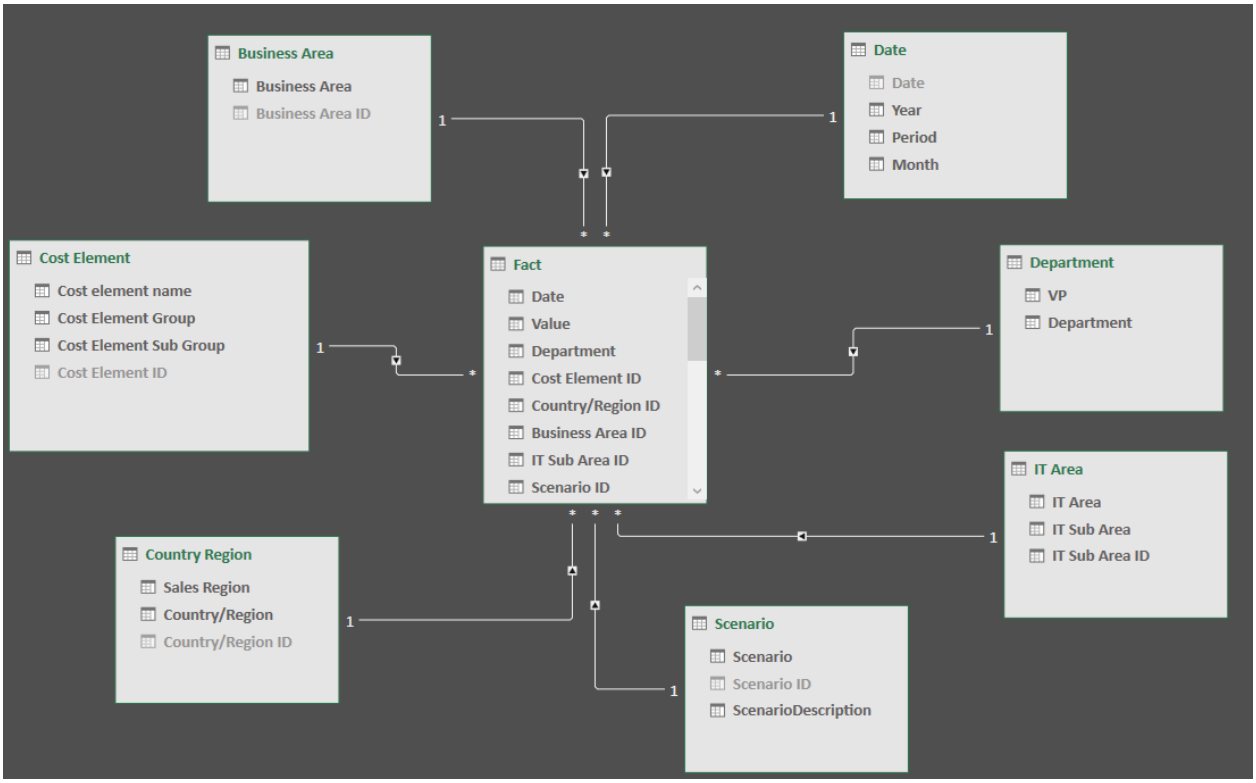


Figure 6.1:

- `diagnose_report(data,output_format ="html")`: https://itsmecevi.github.io/eda-spend-analysis/

with `DataExplorer package`:

- `create_report(data)`: https://itsmecevi.github.io/eda-spend-analysis-2/

Note: Unfortunately, this dataset is a simple and structures dataset. We don't need to change the data or cleaning the parameters.

2. Import the data:

Use this tutorial to import data in Microsoft Power BI: https://bit.ly/2Ikx8QX

3. Create data model:

- Click Relationships tabs (bottom left)
- Let's make the data model from the 8 entity. Just drag every Primary Key of the entity like Figure 6.1.

Data Analysis Expressions:

- **Note:** This use case used DAX (Data Analysis Expressions) for data modeling. More about DAX: https://docs.microsoft.com/en-us/power-bi/desktop-quickstart-learn-dax-basics
- **Extras:**
- Quick Guide to understand the basic of DAX: https://bit.ly/2TPM8Z6
- DAX function reference: https://bit.ly/2GjxzKl

Below are the explanation of DAX `Fact table` in this use case.

Fact DAX attribute:

- `Actual = CALCULATE([Amount], Scenario[ScenarioDescription]="Actual")`
- `Actual/Plan`
  - Value
  - Goal
  - Status
- `Amount = TOTALYTD(SUM([Value]), 'Date'[Date])*.3`
- `LE1 = CALCULATE([Amount], Scenario[ScenarioDescription]="Latest Estimate 1")`
- `LE2 = CALCULATE([Amount], Scenario[ScenarioDescription]="Latest Estimate 2")`
- `LE3 = CALCULATE([Amount], Scenario[ScenarioDescription]="Latest Estimate 3")`
- `Plan = CALCULATE([Amount], Scenario[ScenarioDescription]="Plan")`
- `Var LE1 = [Actual]-[LE1]`
- `Var LE1 % = DIVIDE([Var LE1],[LE1], BLANK())`
- `Var LE2 = [Actual]-[LE2]`
- `Var LE2 % = DIVIDE([Var LE2],[LE2], BLANK())`
- `Var LE3 = [Actual]-[LE3]`
- `Var LE3 % = DIVIDE([Var LE3],[LE3], BLANK())`
- `Var Plan = [Actual]-[Plan]`
- `Var Plan % = DIVIDE([Var Plan],[Plan], BLANK())`

4. Create Report:

KPI Dashboard solution from scratch. In Power BI there are many visualizations, it includes:

- Area charts: Basic (Layered) and Stacked
- Bar and column charts
- Cards: Multi row
- Cards: Single number
- Combo charts
- Doughnut charts
- Funnel charts
- Gauge charts
- Key influencers chart
- KPIs
- Line charts
- Maps: Basic maps
- Maps: ArcGIS maps
- Maps: Filled maps (Choropleth)
- Maps: Shape maps
- Matrix
- Pie charts
- Ribbon chart
- Scatter and Bubble charts
- Scatter-high density
- Slicers
- Tables
- Treemaps
- Waterfall charts

For more info, see the tutorial here: https://bit.ly/2FZwwyA

IT Spend Trend Dashboard: YTD IT Spend Trend Analysis page.

- Var Plan % by Sales Region:

1. Clustered column chart
2. Fields:

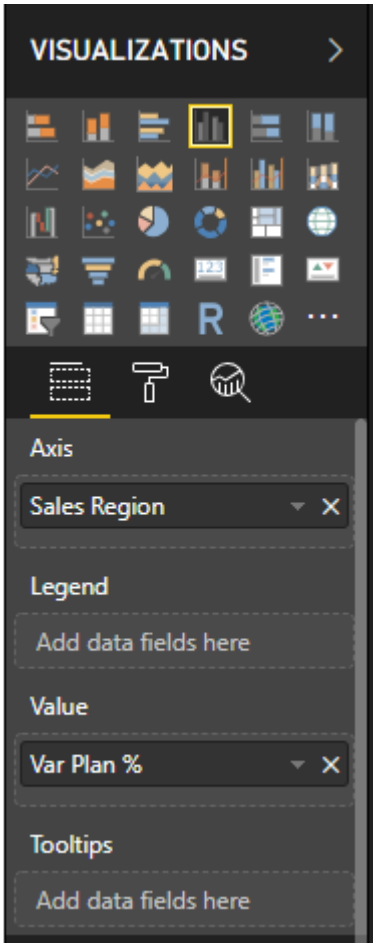


Figure 6.2:

- Country Region: `Sales Region`
- Fact: `Var Plan %`
- See Figure 6.2
- See Figure 6.3

- Var Plan by Month:

1. Line chart
2. Fields:
   - Date: `Month`
   - Fact: `Var Plan %`
   - See Figure 6.4
   - See Figure 6.5

- Var Plan , Var Plan % and Actual by Business Area and Period:

1. Scatter chart
2. Fields:
   - Business Area: `Business Area`
   - Date: `Period`
   - Fact: `Actual`, `Var Plan`, `Var Plan %`
   - See Figure 6.6
   - See Figure 6.7

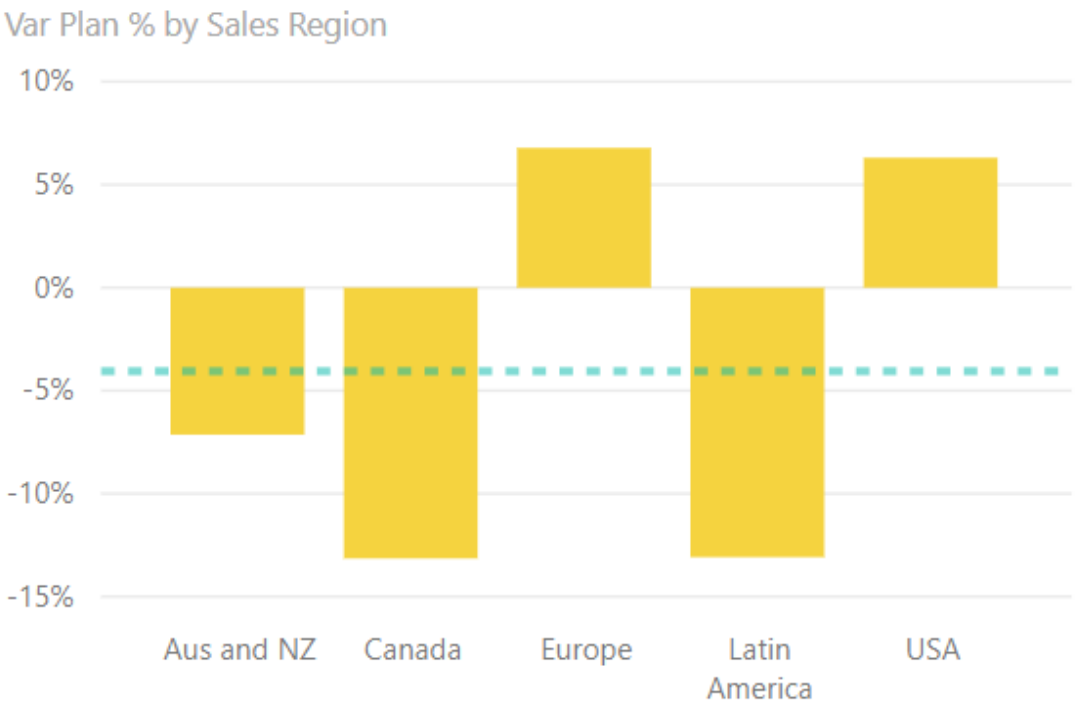


Figure 6.3:

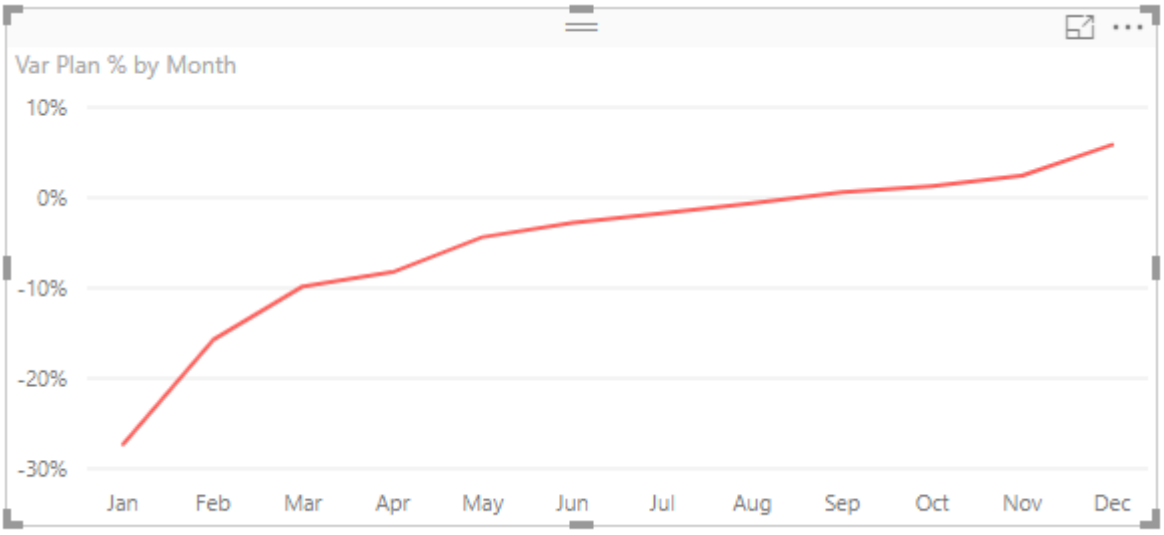


Figure 6.4:

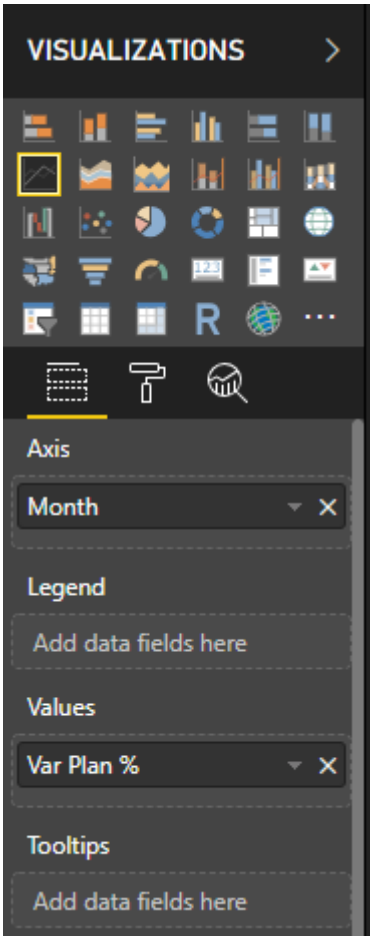


Figure 6.5:

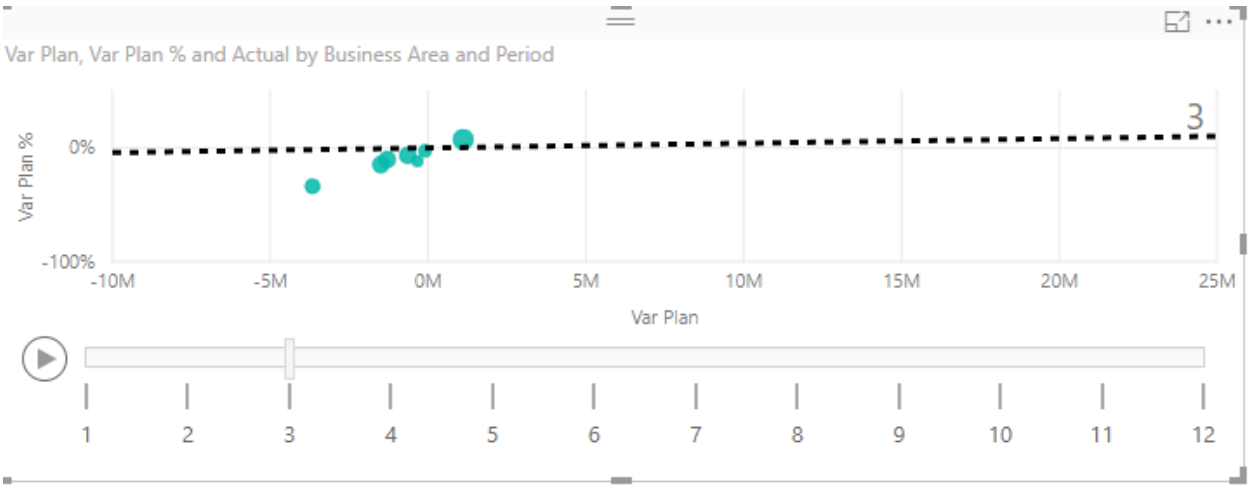


Figure 6.6:

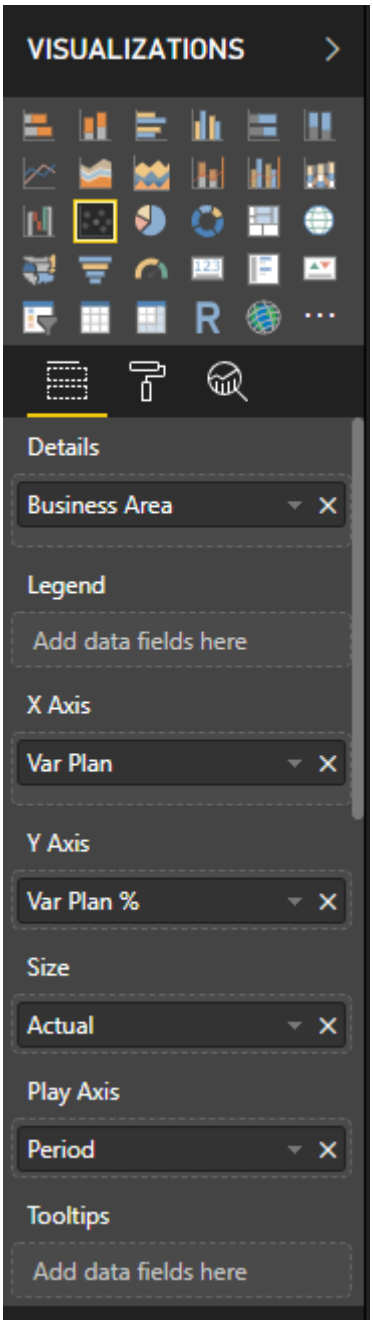

VISUALIZATIONS
Details
Business Area
Legend
Add data fields here
X Axis
Var Plan
Y Axis
Var Plan %
Size
Actual
Play Axis
Period
Tooltips
Add data fields here

Figure 6.7:

Figure 6.8:

- IT Area:

1. Clustered column chart
2. Fields:
   - Fact: `Actual`, `Plan`
   - IT Area: `IT Area`
   - See Figure 6.8
   - See Figure 6.9

Spend by Cost Element Dashboard: YTD Spend by Cost Elements page.

- Var Plan % and Var LE3 % by IT Area:

1. Clustered bar chart
2. Fields:
   - Fact: `Var LE3 %`, `Var Plan %`
   - IT Area: `IT Area`, `IT Sub Area`
   - See Figure 6.12
   - See Figure 6.13

- Var Plan % by Sales Region:

1. Stacked column chart
2. Fields:
   - Country Region: `Sales Region`
   - Fact: `Var Plan %`
   - See Figure 6.14
   - See Figure 6.15

- Amount by Month and Scenario:

1. Clustered bar chart
2. Fields:
   - Date: `Month`
   - Fact: `Amount`
   - Scenario: `Scenario`
   - See Figure 6.16
   - See Figure 6.17

Plan Variance Analysis Dashboard.

- Variance Latest Estimates:

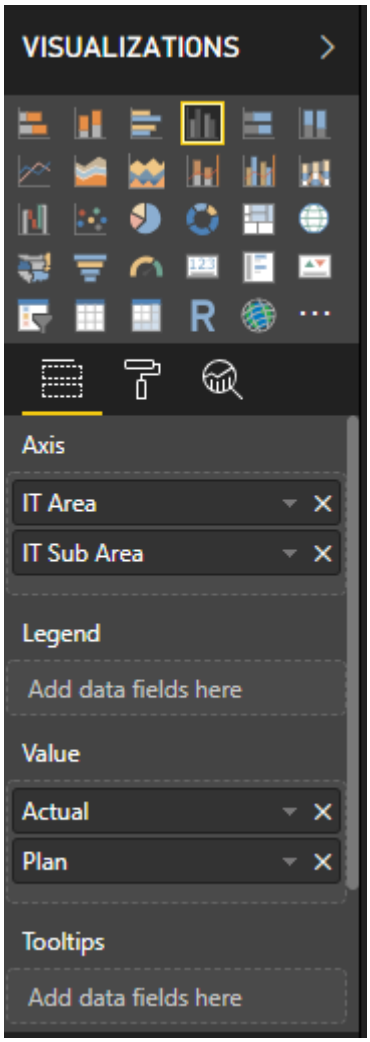


Figure 6.9:

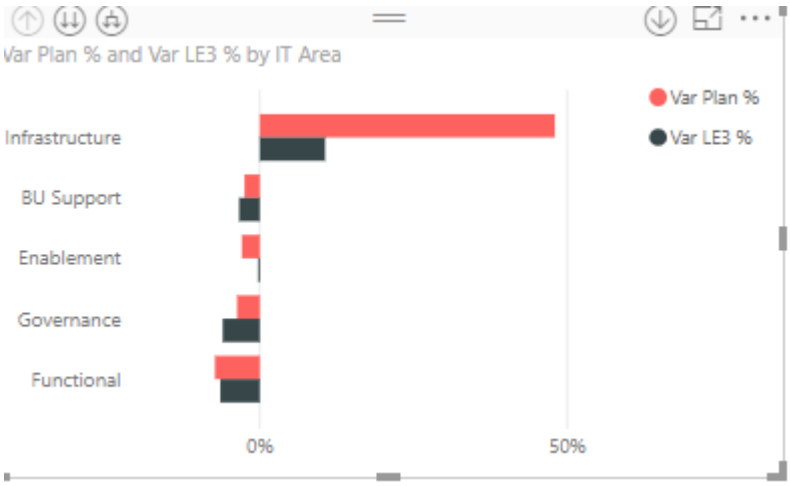


Figure 6.10:

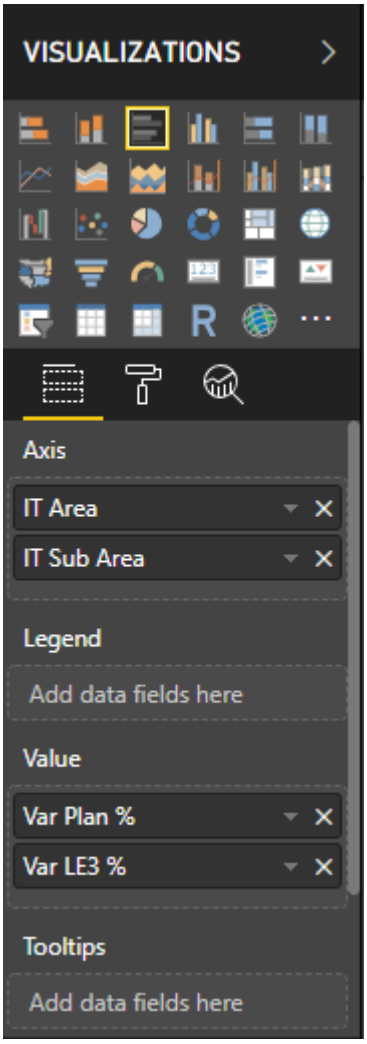

VISUALIZATIONS
Axis
IT Area
IT Sub Area
Legend
Add data fields here
Value
Var Plan %
Var LE3 %
Tooltips
Add data fields here

Figure 6.11:

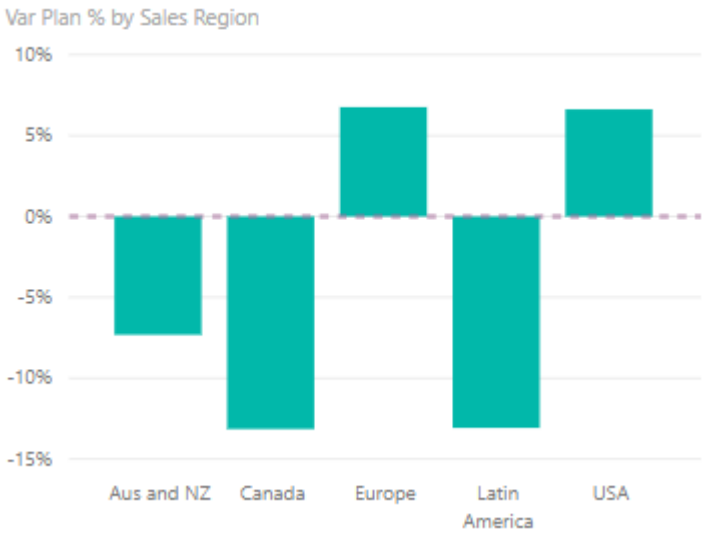


Figure 6.12:

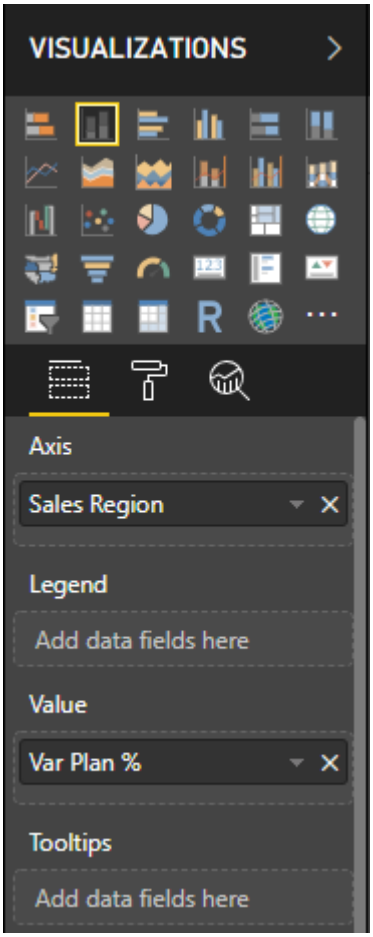


Figure 6.13:

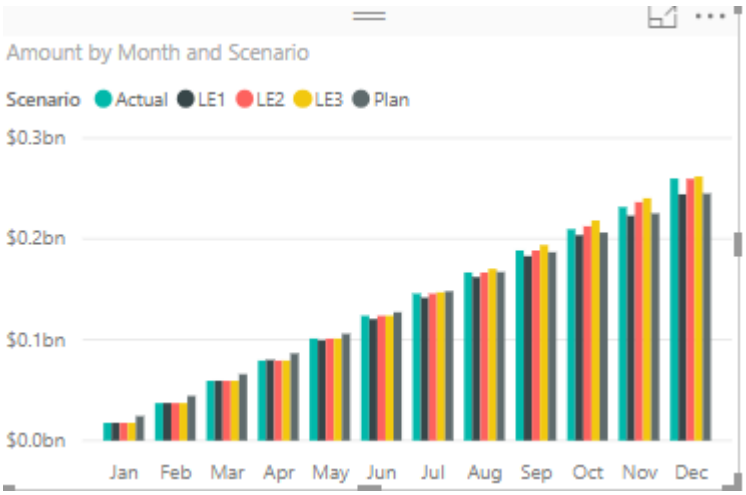


Figure 6.14:

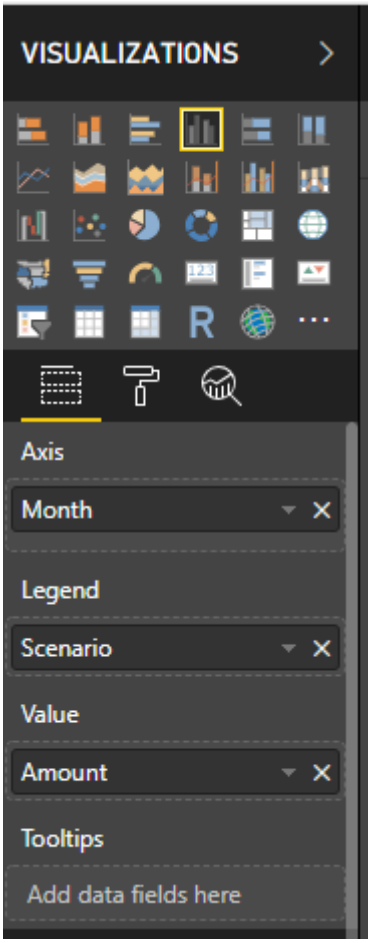


Figure 6.15:

Figure 6.16:

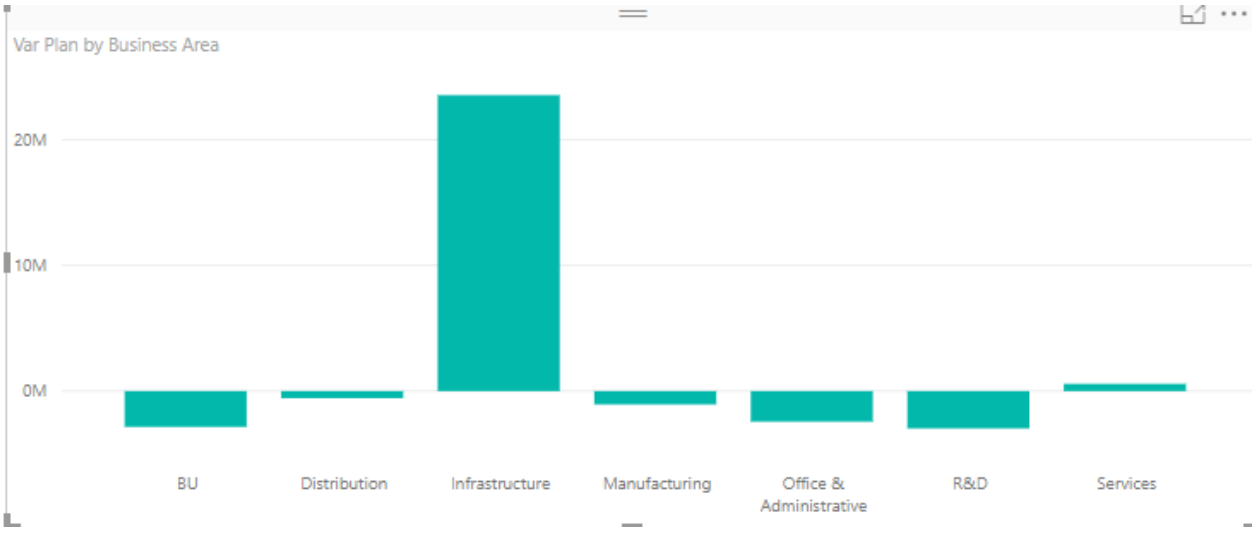


Figure 6.17:

1. Multi-row card
2. Fields:
   - Fact: `Var Plan %`
   - Fact: `Var LE1 %`
   - Fact: `Var LE2 %`
   - Fact: `Var LE3 %`
   - See Figure 6.18

- Var Plan by Business Area:

1. Clustered column chart
2. Fields:
   - Business Area: `Business Area`
   - Fact: `Var Plan`
   - See Figure 6.19
   - See Figure 6.20

- Var Plan % by Business Area:

1. Clustered column chart
2. Fields:
   - Business Area: `Business Area`
   - Fact: `Var Plan`
   - See Figure 6.21
   - See Figure 6.22

- Var Plan by Sales Region and country :

1. Clustered column chart

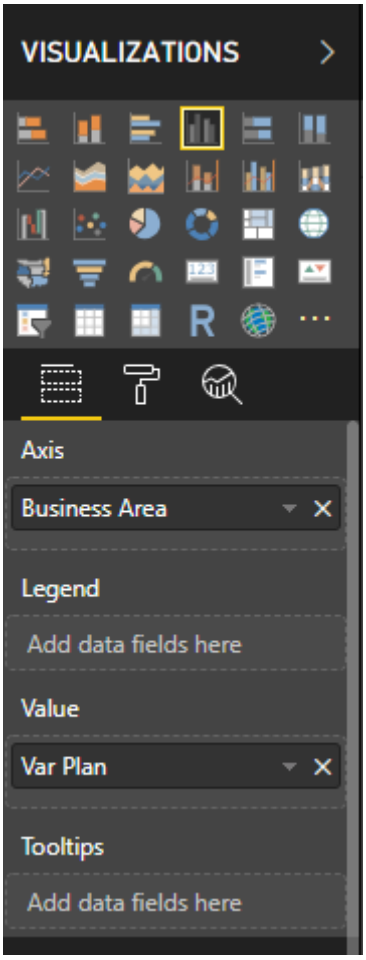


Figure 6.18:

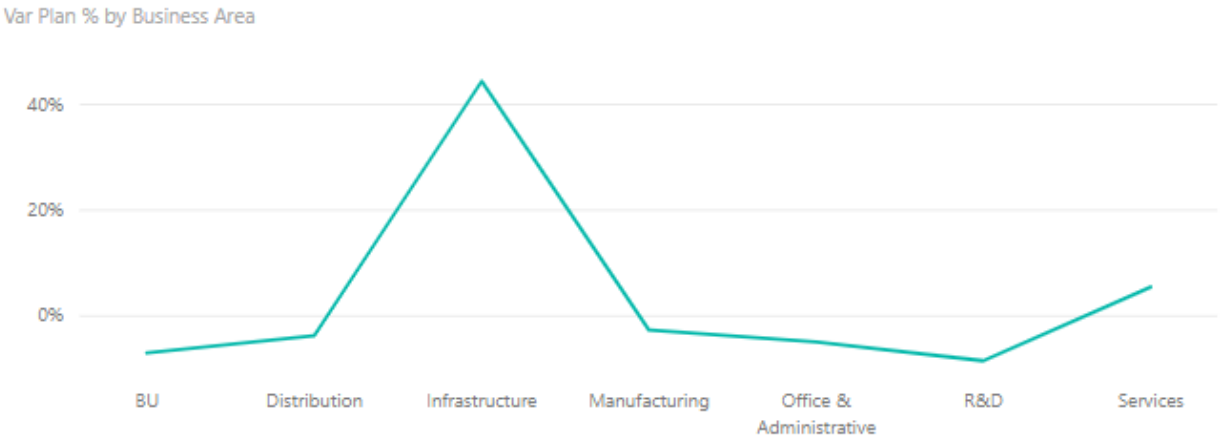


Figure 6.19:

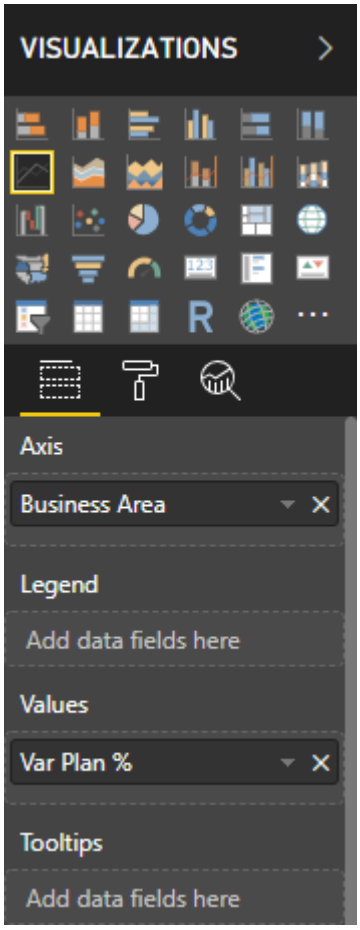

VISUALIZATIONS
Axis
Business Area
Legend
Add data fields here
Values
Var Plan %
Tooltips
Add data fields here

Figure 6.20:

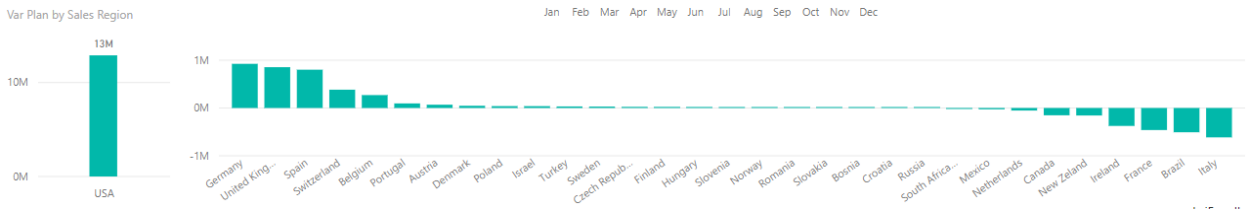


Figure 6.21:

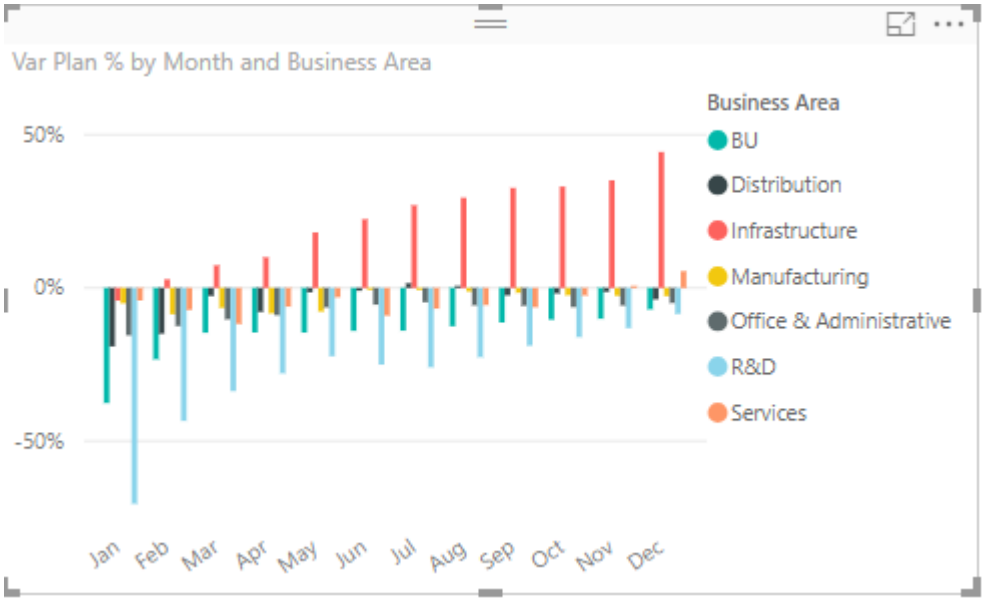


Figure 6.22:

2. Fields:
   - Country Region: `Sales Region`
   - Fact: `Var Plan`
   - See Figure 6.23

- Var Plan % by Month and Business Area:

1. Clustered column chart
2. Fields:
   - Business Area: `Business Area`
   - Date: `Month`
   - Fact: `Var Plan %`
   - See Figure 6.24
   - See Figure 6.25

6.1.6 Sources

- `Spend Analysis 101: How, Why, and What To Do with the Data [70]`
- `Spend Analysis 101 [71]`
- `IT Spend Analysis sample for Power BI [72]`
- `Add visualizations to a Power BI report [73]`
- `DAX function reference [74]`

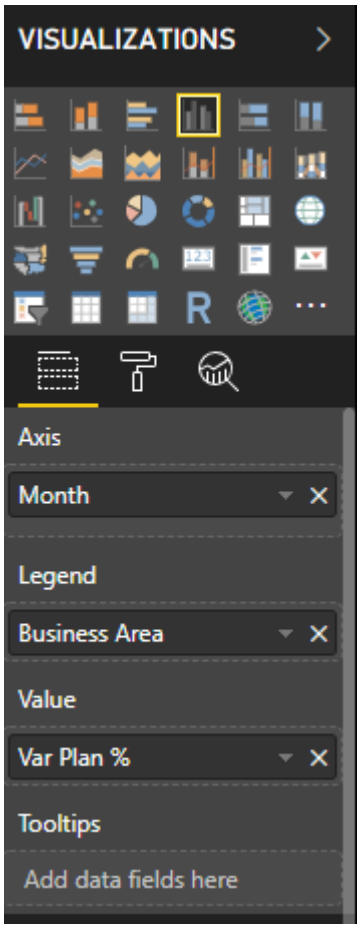

VISUALIZATIONS
Axis
Month
Legend
Business Area
Value
Var Plan %
Tooltips
Add data fields here

Figure 6.23:

6.2 Human Resources KPI

6.2.1 Overview

Human Resources (HR) dashboards aggregate and present employee data in a meaningful way. As a metrics, Human Resources dashboards simplify information gathering, and present data in a way that can be sorted, analyzed, and presented. This use case dashboard looks at new hires, active employees, and employees who left and tries to uncover any trends in the hiring strategy. Our main objectives are to understand:

- Who we hire
- Biases in our hiring strategy (over budget, time, low performances employee)
- Trends in employees hiring process (stereotype, gender, region, etc)

6.2.2 Dataset

This Human Resources data is part of a series that illustrates how you can Dashboard BI with business-oriented data, reports and dashboards. This is real data from obviEnce (http://obvience.com/) that has been anonymized. Find the data in my dropbox (too big in github repository): https://www.dropbox.com/s/u70jscu18ruqk49/Dataset-Human%20Resources%20Sample.xlsx?dl=0

This dataset contains 9 entity (table) and every table has attribute to.

1. age group

- Attribute: AgeGroupID, AgeGroup

2. gender

- Attribute: ID, Gender, Sort

3. ethnicity Attribute: Ethnic Group, Ethnicity
4. date Attribute: Date, Month, MonthNumber, Period, PeriodNumber, Qtr, QtrNumber, Year, Day, MonthStartDate, MonthEndDate, MonthIncrementNumber
5. separation reason Attribute: SeparationTypeID, SeparationReason
6. pay type Attribute: PayTypeID, PayType
7. FP

- Attribute: FP, FPDesc

8. bu Attribute: BU, RegionSeq, VP, Region

- employee_fact Attribute: date, EmplID, Gender, Age, EthnicGroup, FP, TermDate, isNewHire, BU, HireDate, PayTypeID, TermReason, AgeGroupID ,TenureDays, TenureMonths, BadHires

6.2.3 Glosarry

SPLY: same periode last year. One year back in time from the context dates YoY: Year-over-year (YOY) is a method of evaluating two or more measured events to compare the results at one period with those of a comparable period on an annualized basis Seps: Separation of Employees FPDes: Full-Time or Part-Time Employees Description BU: Business Unit VP: Vice Predident

6.2.4 Solution

New hires Dashboard:

- New Hires Vs New Hires SPLY
- New Hires by Region and FPDesc
- New Hires by Region and Ethnicity
- Gender
- AgeGroup

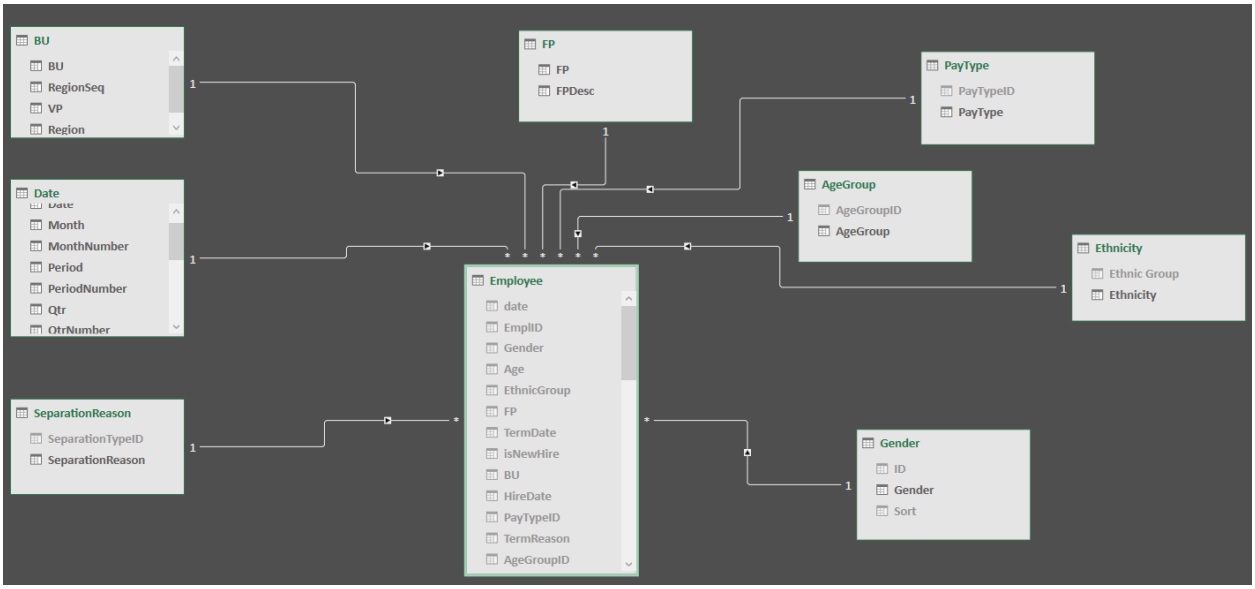


Figure 6.24:

Active Employees vs Separations Dashboard:

- YoY Change
- AgeG roup
- Gender
- Actives by Region
- Actives vs Actives SPLY
- Separations by Reason
- Seps vs Seps SPLY

Bad Hires Dashboard: Bad hires are defined as employees who didn't last for more than 60 days.

- Bad Hires by Region and Gender
- Bad Hires by Region and Ethnicity
- Bad Hires YoY % Change by Month and Age Group
- Bad Hire % of Actives and Bad Hires % of Active SPLY by Region

6.2.5 Technical Documentation

Below are the steps by step to make KPI Dashboard for this solution.

1. Exploratory Data Analysis with R Programming:

See the section `3.2 EDA: Exploratory Data Analysis` for doing the data cleaning. The dataset based on `Star Schema`. Used `dlookr package` and `DataExplorer package`. The data isn't messy data. We can easily going further.

2. Import the data:

Use this tutorial to import data in Microsoft Power BI: https://bit.ly/2Ikx8QX

3. Create data model:

- Click Relationships tabs (bottom left)
- Let's make the data model from the 8 entity. Just drag every Primary Key of the entity like Figure 6.26.

Data Analysis Expressions:

- **Note:** This use case used DAX (Data Analysis Expressions) for data modeling. More about DAX: https://docs.microsoft.com/en-us/power-bi/desktop-quickstart-learn-dax-basics
- **Extras:**
- Quick Guide to understand the basic of DAX: https://bit.ly/2TPM8Z6
- DAX function reference: https://bit.ly/2GjxzKl

DAX of `Fact table` (See the section 2 for the data models: `Stage Four - Data Modeling`).

Employee DAX attribute:

- `Actives = CALCULATE([EmpCount], FILTER(Employee, ISBLANK(Employee[TermDate])))`
- `Actives SPLY = CALCULATE([Actives],SAMEPERIODLASTYEAR('Date'[Date]))`
- `Actives YoY % Change = DIVIDE([Actives YoY Var], [Actives SPLY])`
- `Actives YoY Var = [Actives]-[Actives SPLY]`
- `AVG Age = ROUND(AVERAGE([Age]), 0)`
- `AVG Tenure Days = AVERAGE([TenureDays])`
- `AVG Tenure Months = ROUND([AVG Tenure Days]/30, 1)-1`
- `Bad Hires SPLY = CALCULATE([Sum of BadHires],SAMEPERIODLASTYEAR('Date'[Date]))`
- `Bad Hires YoY % Change = DIVIDE([Bad Hires YoY Var], [Bad Hires SPLY])`
- `Bad Hires YoY Var = [Sum of BadHires]-[Bad Hires SPLY]`
- `BadHire%ofActives = DIVIDE([Sum of BadHires],[Actives])`
- `BadHire%ofActiveSPLY = DIVIDE([Bad Hires SPLY],[Actives SPLY])`
- `BadHires = IF(OR(((HireDate]-[TermDate])*-1)>=61,ISBLANK([TermDate])),0,1)`
- `EmpCount = CALCULATE(COUNT([EmplID]), FILTER(ALL('Date'[PeriodNumber]), 'Date'[PeriodNumber] = MAX('Date'[PeriodNumber])))`
- `EmpCount SPLY = CALCULATE(COUNT([EmplID]), FILTER(ALL('Date'[PeriodNumber]), 'Date'[PeriodNumber] = MAX('Date'[PeriodNumber])),SAMEPERIODLASTYEAR('Date'[Date]))`
- `New Hires = SUM([isNewHire])`
- `New Hires SPLY = CALCULATE([New Hires],SAMEPERIODLASTYEAR('Date'[Date]))`
- `New Hires YoY % Change = DIVIDE([New Hires YoY Var], [New Hires SPLY])`
- `New Hires YoY Var = [New Hires]-[New Hires SPLY]`
- `Sep%ofActive = DIVIDE([Seps],[Actives])`
- `Sep%ofSMLYActives = DIVIDE([Seps SPLY],[Actives SPLY])`
- `Seps = CALCULATE(COUNT([EmplID]), FILTER(Employee, NOT(ISBLANK(Employee[TermDate]))))`
- `Seps SPLY = CALCULATE([Seps],SAMEPERIODLASTYEAR('Date'[Date]))`
- `Seps YoY % Change = DIVIDE([Seps YoY Var], [Seps SPLY])`
- `Seps YoY Var = [Seps]-[Seps SPLY]`
- `Sum of BadHires = SUM([BadHires])`
- `TO % = DIVIDE([Seps], [Actives])`
- `TO % Norm = CALCULATE([TO %], all(Gender[Gender]), ALL(Ethnicity[Ethnicity]))`
- `TO % Var = [TO %]-[TO % Norm]`

4. Create Report:

KPI Dashboard solution from scratch. In Power Bi there are many area to create visualization, it include:

- Area charts: Basic (Layered) and Stacked
- Bar and column charts
- Cards: Multi row
- Cards: Single number
- Combo charts
- Doughnut charts
- Funnel charts
- Gauge charts

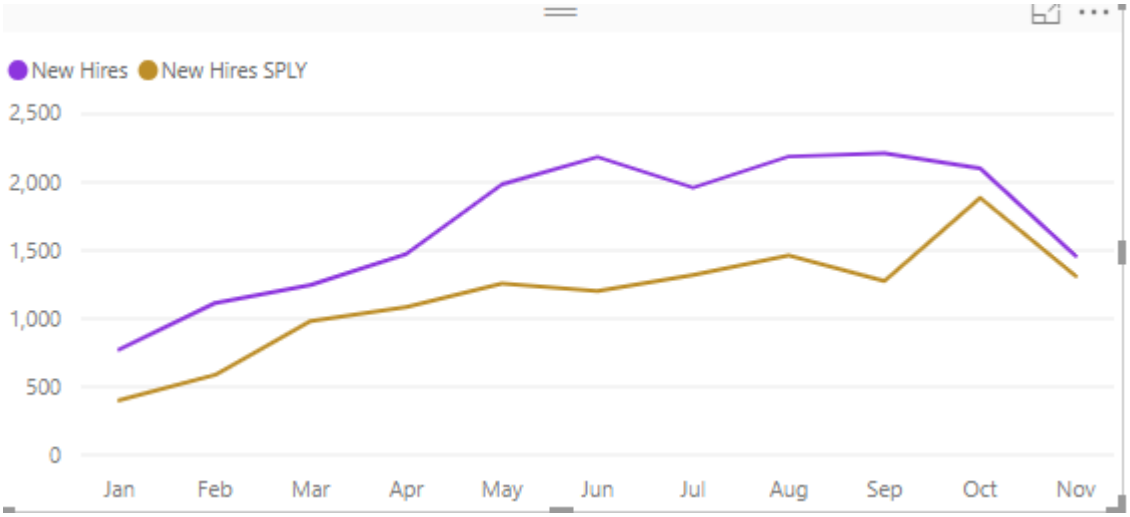


Figure 6.25:

- Key influencers chart
- KPIs
- Line charts
- Maps: Basic maps
- Maps: ArcGIS maps
- Maps: Filled maps (Choropleth)
- Maps: Shape maps
- Matrix
- Pie charts
- Ribbon chart
- Scatter and Bubble charts
- Scatter-high density
- Slicers
- Tables
- Treemaps
- Waterfall charts

For more info, see the tutorial here: https://bit.ly/2FZwwyA

New hires Dashboard.

- New Hires Vs New Hires SPLY (Same Periode Last Year):

1. Line chart
2. Fields:
   - Date: `Month`
   - Employee: `New Hires`, `New Hires SPLY`
   - See Figure 6.27

- New Hires by Region and FPDesc:

1. Stacked column chart
2. Fields:
   - BU: `Region`, `VP`
   - Employee: `New Hires`
   - FP: `FPDesc`
   - See Figure 6.28

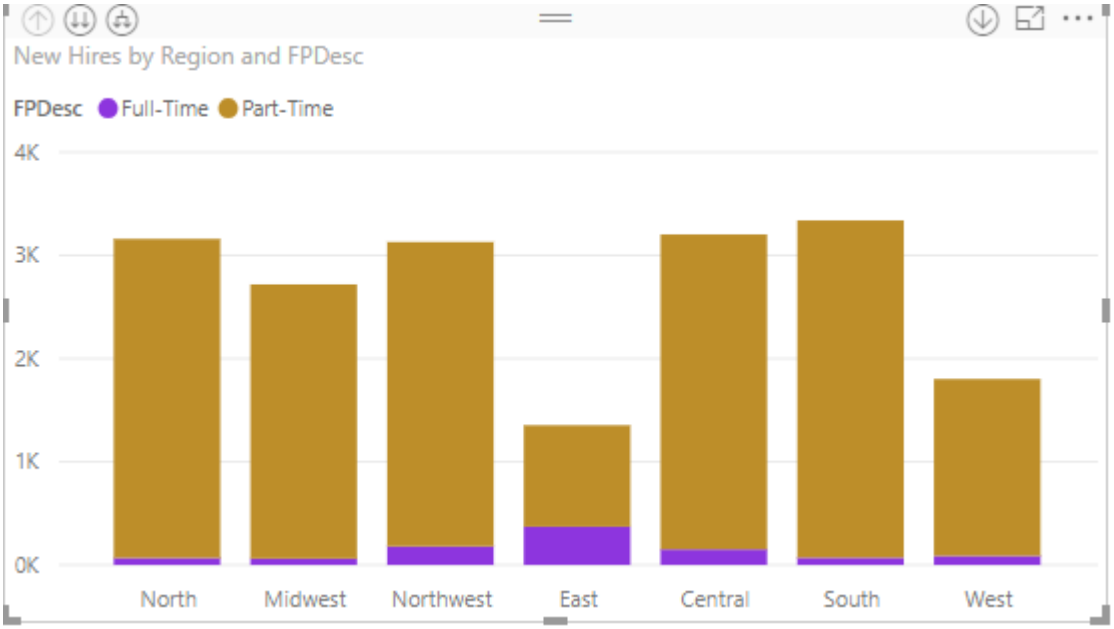


Figure 6.26:

- New Hires by Region and Ethnicity:

1. Stacked column chart
2. Fields:
   - BU: `Region`, `VP`
   - Employee: `New Hires`
   - Ethnicity: `Ethnicity`
   - See Figure 6.29

- Gender:

1. Pice chart
2. Fields:
   - Employee: `New Hires`
   - Gender: `Gender`
   - See Figure 6.30

- AgeGroup:

1. Pice chart
2. Fields:
   - AgeGroup: `AgeGroup`
   - Employee: `New Hires`
   - See Figure 6.31

Active Employees vs Separations Dashboard.

- YoY Change:

1. Multi-row card
2. Fields:
   - Date: `Month`
   - Employee: `Actives YoY % Change`, `Seps YoY % Change`
   - See Figure 6.32

Figure 6.27:

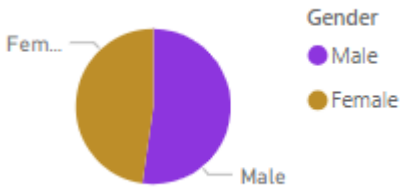


Figure 6.28:

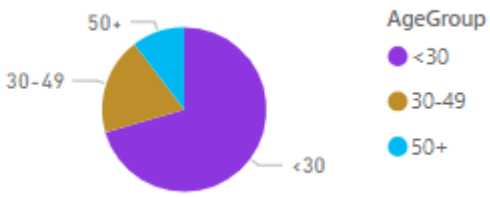


Figure 6.29:


Jan
18.6%
Seps YoY % Change
3.2%
Actives YoY % Change
Feb
22.0%
Seps YoY % Change
4.4%
Actives YoY % Change
Mar
24.2%
Seps YoY % Change
4.5%
Actives YoY % Change
Apr
26.9%
Seps YoY % Change
5.0%
Actives YoY % Change
May
31.5%
Seps YoY % Change

Figure 6.30:

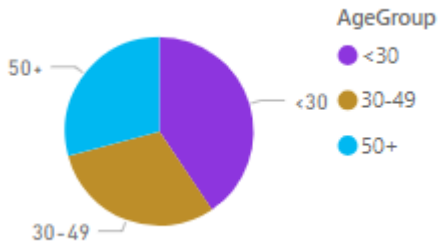


Figure 6.31:

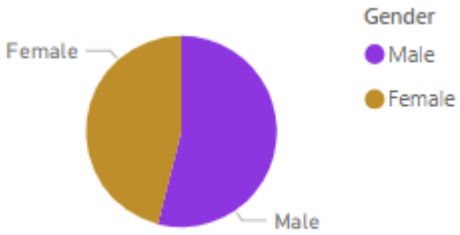


Figure 6.32:

- AgeGroup:

1. Pie chart
2. Fields:
   - AgeGroup: `AgeGroup`
   - Employee: `Actives`
   - See Figure 6.33

- Gender:

1. Pie chart
2. Fields:
   - Gender: `Gender`
   - Employee: `Actives`
   - See Figure 6.34

- Actives by Region:

1. Clustered bar chart
2. Fields:
   - BU: `Region`, `VP`
   - Employee: `Actives`
   - See Figure 6.35

- Actives vs Actives SPLY:

1. Clustered column chart
2. Fields:
   - BU: `Region`, `VP`
   - Date: `Month`
   - Employee: `Actives`, `Actives SPLY`

Figure 6.33:

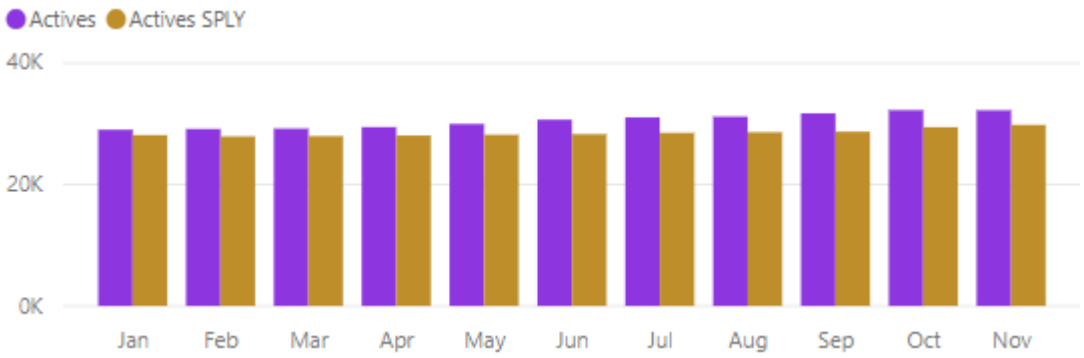


Figure 6.34:

- See Figure 6.36

- Separations by Reason:

1. Stacked bar chart
2. Fields:
   - Employee: `Seps`
   - SeparationReason: `SeparationReason`
   - See Figure 6.37

- Seps vs Seps SPLY:

1. Clustered column chart
2. Fields:
   - BU: `Region`, `VP`
   - Date: `Month`
   - Employee: `Seps`, `Seps SPLY`
   - See Figure 6.38

Bad Hires Dashboard: Bad hires are defined as employees who didn't last for more than 60 days.

Bad Hires (<60 day employment).

- Bad Hires by Region and Gender:

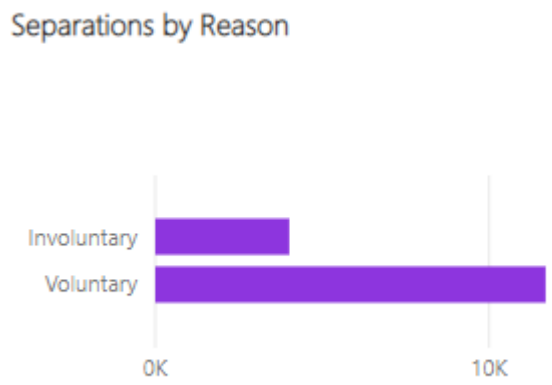


Figure 6.35:

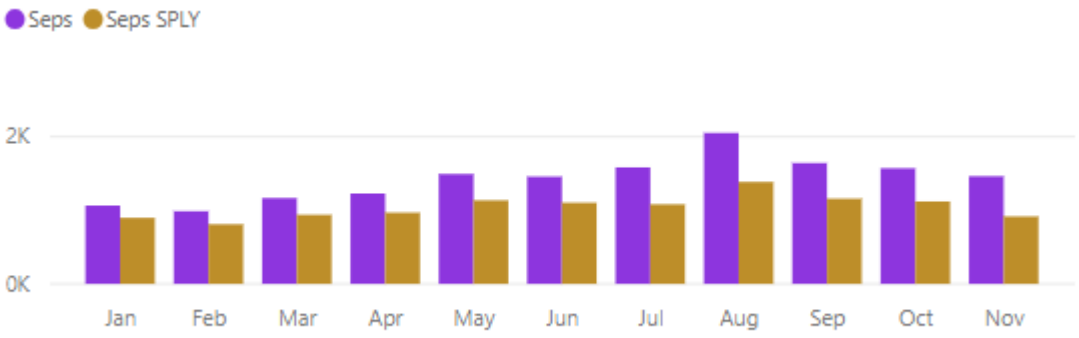


Figure 6.36:

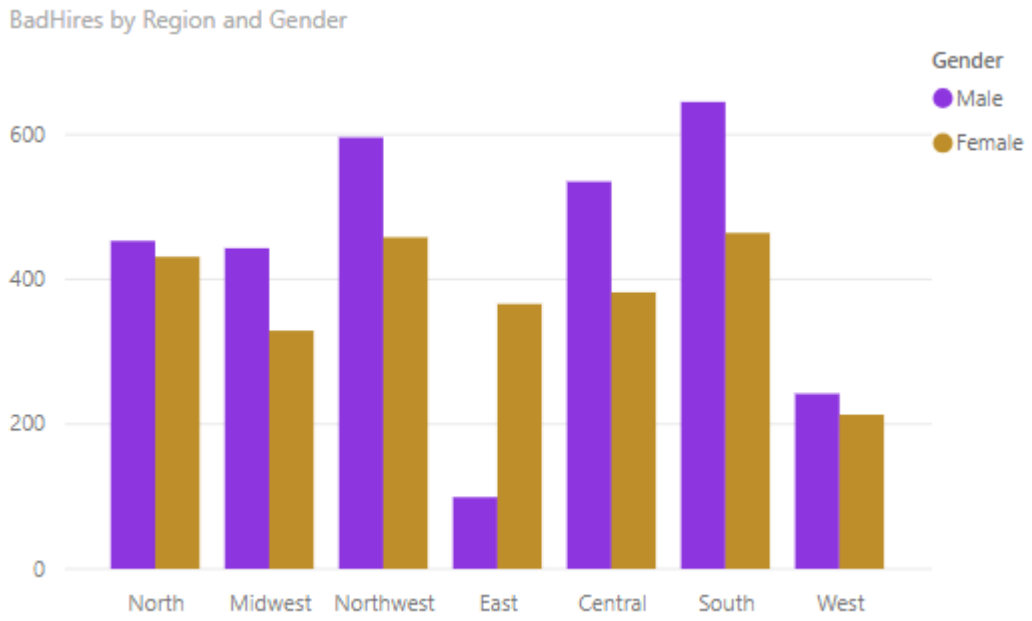


Figure 6.37:

1. Clustered column chart
2. Fields:
   - BU: `Region`, `VP`
   - Employee: `Bad Hires`
   - Gender: `Gender`
   - See Figure 6.39

- Bad Hires by Region and Ethnicity:

1. 100% Stacked column chart
2. Fields:
   - BU: `Region`, `VP`
   - Employee: `Bad Hires`
   - Ethnicity: `Ethnicity`
   - See Figure 6.40

- Bad Hires YoY % Change by Month and Age Group:

1. Clustered column chart
2. Fields:
   - AgeGroup: `AgeGroup`
   - BU: `Region`, `VP`
   - Date: `Month`
   - Employee: `Bad Hires YoY % Change`
   - See Figure 6.41

- Bad Hire % of Actives and Bad Hires % of Active SPLY by Region:

1. Clustered column chart
2. Fields:
   - BU: `Region`, `VP`
   - Employee: `BadHire%ofActives`, `BadHire%ofActiveSPLY`
   - See Figure 6.42

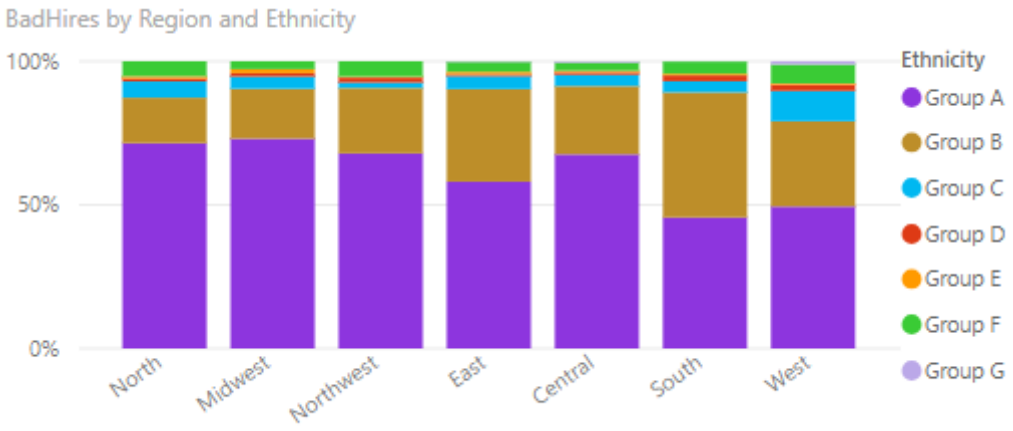


Figure 6.38:

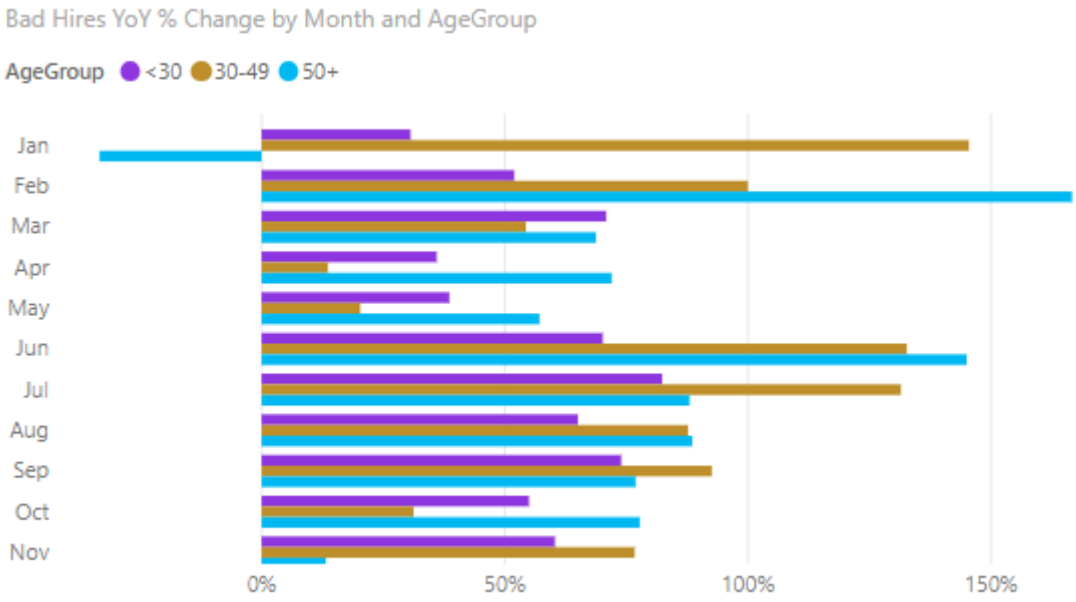


Figure 6.39:

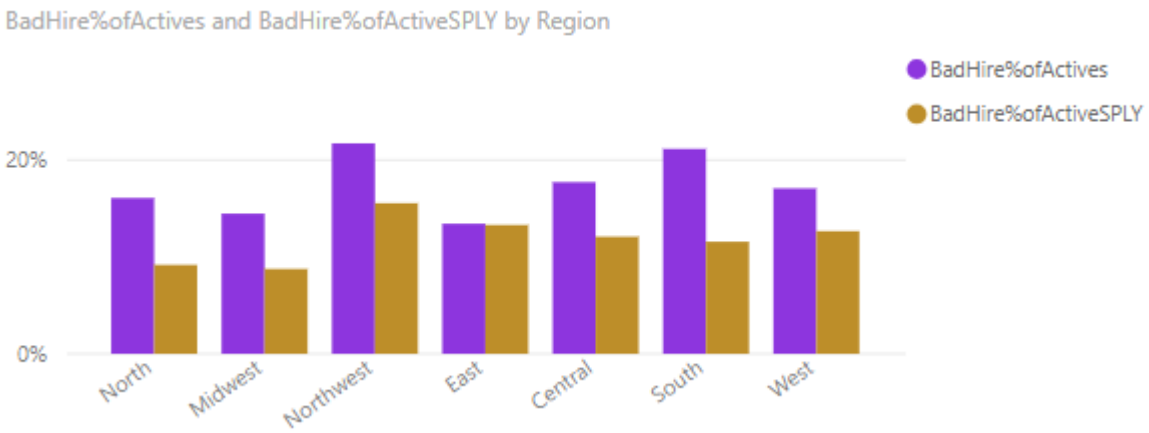


Figure 6.40:

6.2.6 Sources

- `Human Resources sample for Power BI [75]`

6.3 Sales Opportunity Analysis KPI

6.3.1 Overview

Opportunity analyisis is a detailed review of the prospects within a potential market. The company has 2 sales channels: direct and partner. The Sales Department created this dashboard to track opportunities and revenue by region, deal size, and channel.

The Sales Department relies on two measures of revenue:

- **Revenue** – this is a salesperson's estimate of what he believes the revenue will be.
- **Factored Revenue** – this is calculated as Revenue X Probability% and is generally accepted as being a more-accurate predictor of actual sales revenue. Probability is determined by the deal's current **Sales Stage** .
  - Lead – 10%
  - Qualify – 20%
  - Solution – 40%
  - Proposal – 60%
  - Finalize – 80%

6.3.2 Dataset

This data is part of a series that illustrates how you can Dashboard BI with business-oriented data, reports and dashboards. This is real data from obviEnce (http://obvience.com/) that has been anonymized. Find the dataset in my github repository: https://github.com/itsmecevi/opportunity-sales-dataset

This dataset contains 5 entity (table) and every table has attribute.

1. sales stage

- Attibute: Probability, Sales Stage, Sales Stage ID

2. product

- Attibute: Product Code, Product ID

3. partner

- Attibute: Partner, Partner ID, Partner Driven

4. opportunity

- Attibute: Name, Opportunity ID, Rank, SizeID, Opportunity Size

5. fact

- Attibute: EstimatedCloseDate, Opportunity ID, Sales Stage ID, Account ID, Partner ID, Product ID, ProductRevenue, FactoredProductRevenue, Create Date, Opportunity Days, Year, Month_Number, Month

6.3.3 Glossary

AVG: Average

6.3.4 Solution

Opportunity Counts Overview:

- Opportunity Count
- Opportunity Count by Region
- Opportunity Count by Sales Stage
- Opportunity Count by Partner Driven and Opportunity Size
- Opportunity Count by Partner Driven and Sales Stage

Revenue Analysis:

- Revenue by Region
- Revenue by Sales Stage and Partner Driven
- Revenue
- Factored Revenue
- Opportunity Count
- AVG Revenue by Partner Driven and Opportunity Size

Upcoming Opportunities by Month:

- Opportunity Count
- Factored Revenue by Opportunity Size
- AVG Revenue by Partner Driven and Sales Stage
- Opportunity Count by Month and Sales Stage

6.3.5 Technical Documentation

Below are the steps by step to make Sales Opportunity KPI Dashboard for this solution.

1. Exploratory Data Analysis with R Programming:

See the section `3.2 EDA: Exploratory Data Analysis` for doing the data cleaning. The dataset based on `Star Schema`. Used `dlookr package` and `DataExplorer package`. The data isn‘t messy data. We can easily going further.

2. Import the data:

Import:

- In Power BI, click Get Data in the lower left screen.
- Under Import or Connect to Data > Files, click Get.
- Click Local File.
- Choose which file to upload and click Open.

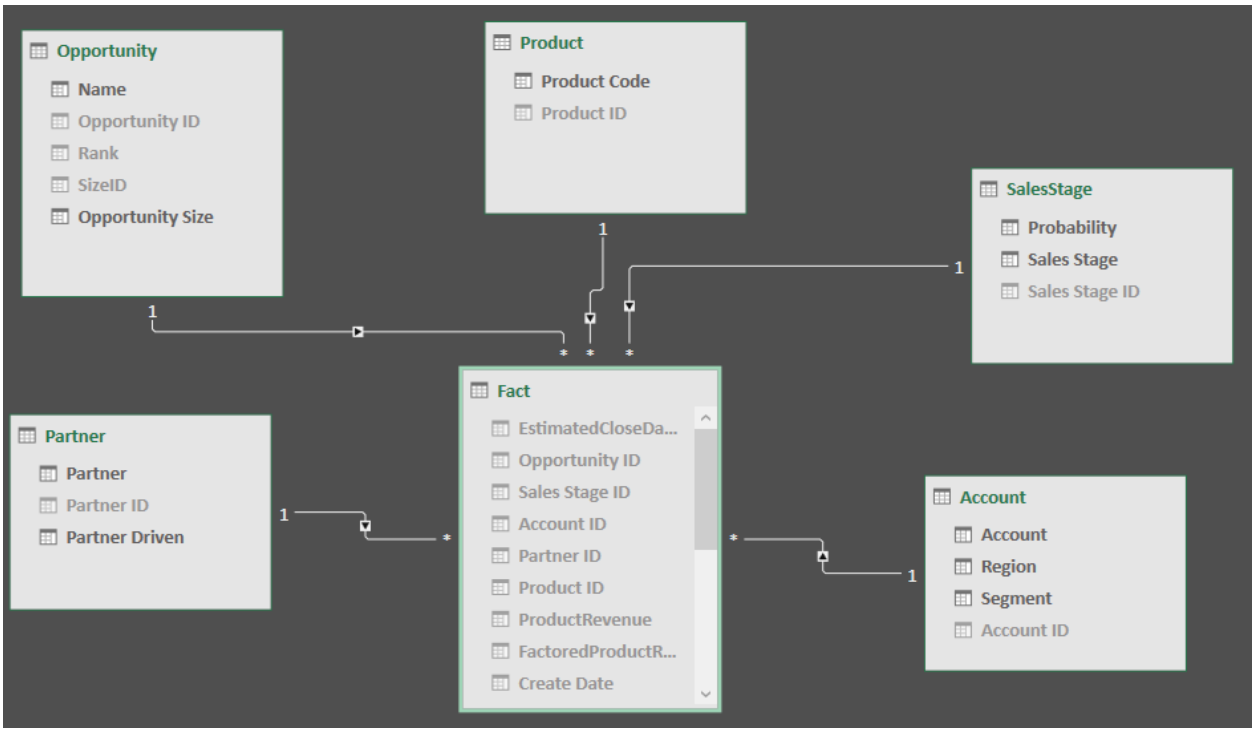


Figure 6.41:

- Click Upload under Upload your Excel file to Power BI.
- The message "Your file has been uploaded" should appear.

3. Create data model:

- Click Relationships tabs (bottom left)
- Let's make the data model from the 5 entity. Just drag every Primary Key of the entity like Figure 6.43

Data Analysis Expressions:

- **Note:** This use case used DAX (Data Analysis Expressions) for data modeling. More about DAX: https://docs.microsoft.com/en-us/power-bi/desktop-quickstart-learn-dax-basics
- **Extras:**
- Quick Guide to understand the basic of DAX: https://bit.ly/2TPM8Z6
- DAX function reference: https://bit.ly/2GjxzKl

DAX of `Fact table` of this dataset (See the section 2 for the data models: `Stage Four - Data Modeling`).

- `Avg Opportunity Days = AVERAGE([Opportunity Days])`
- `Avg Revenue = AVERAGE([ProductRevenue])`
- `Factored Revenue = SUM([FactoredProductRevenue])`
- `Month = FORMAT([EstimatedCloseDate],"MMM")`
- `Opportunity Count = COUNTA([Opportunity ID])`
- `Revenue = SUM([ProductRevenue])`
- `Tot Opportunity Days = sum([Opportunity Days])`
- `Year = Year([EstimatedCloseDate])`

4. Create Report:

Figure 6.42:

KPI Dashboard solution from scratch. In Power Bi there are many area to create visualization, it include:

- Area charts: Basic (Layered) and Stacked
- Bar and column charts
- Cards: Multi row
- Cards: Single number
- Combo charts
- Doughnut charts
- Funnel charts
- Gauge charts
- Key influencers chart
- KPIs
- Line charts
- Maps: Basic maps
- Maps: ArcGIS maps
- Maps: Filled maps (Choropleth)
- Maps: Shape maps
- Matrix
- Pie charts
- Ribbon chart
- Scatter and Bubble charts
- Scatter-high density
- Slicers
- Tables
- Treemaps
- Waterfall charts

For more info, see the tutorial here: https://bit.ly/2FZwwyA

Opportunity Counts Dashboard

- Opportunity Count:

1. Multi-row card
2. Fields:
   - Fact: `Opportunity Count`
   - See Figure 6.44

- Opportunity Count by Region:

1. Pie Chart
2. Fields:
   - Account: `Region`
   - Fact: `Opportunity Count`
   - See Figure 6.45

- Opportunity Count by Sales Stage:

1. Stacked bar chart

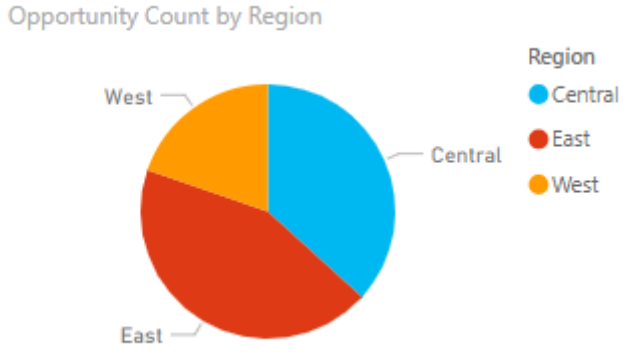


Figure 6.43:

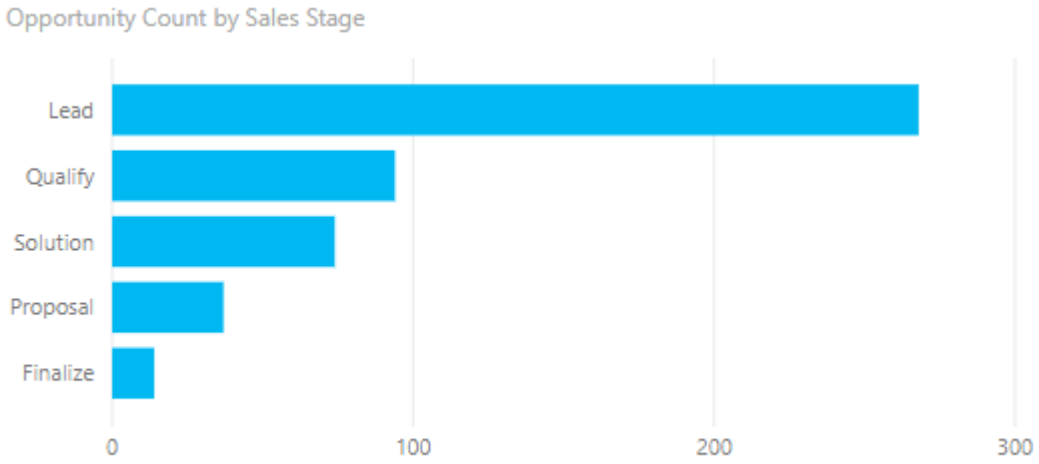


Figure 6.44:

2. Fields:
   - Fact: `Opportunity Count`
   - SalesStage: `Sales Stage`
   - See Figure 6.46

- Opportunity Count by Partner Driven and Opportunity Size:

1. Clustered column chart
2. Fields:
   - Fact: `Opportunity Count`
   - Opportunity: `Opportunity Size`
   - Partner: `Partner Driven`
   - See Figure 6.47

- Opportunity Count by Partner Driven and Sales Stage:

1. Clustered column chart
2. Fields:
   - Fact: `Opportunity Count`

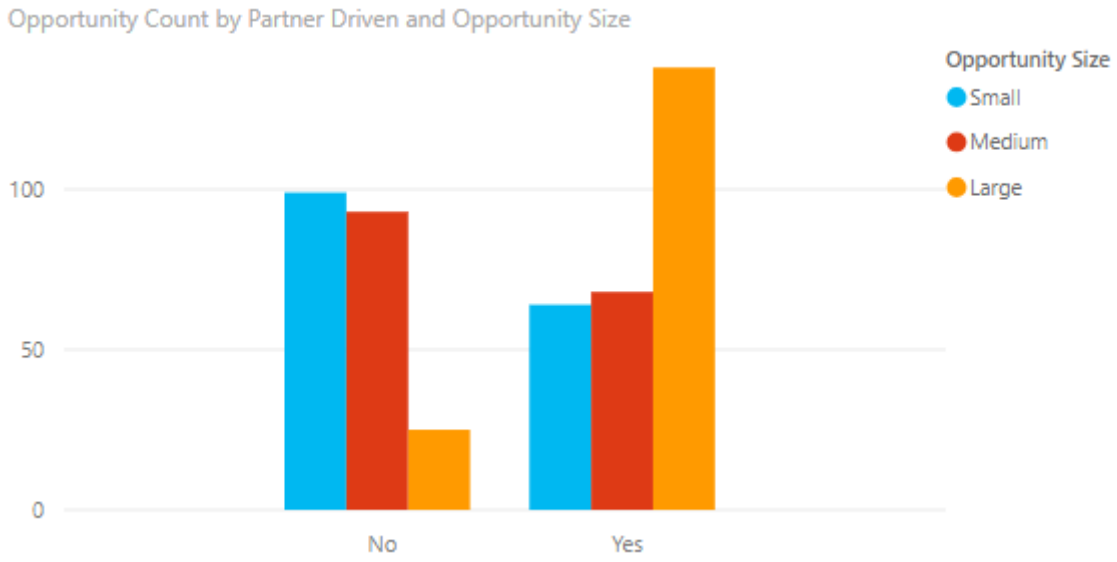


Figure 6.45:

- Partner: `Partner Driven`
- SalesStage: `Sales Stage`
- See Figure 6.48

Revenue Analysis Dashboard

- Revenue by Region:

1. Pie chart
2. Fields:
   - Account: `Region`
   - Fact: `Revenue`
   - See Figure 6.49

- Revenue by Sales Stage and Partner Driven:

1. Clustered column chart
2. Fields:
   - Fact: `Revenue`
   - Partner: `Partner Driven`
   - SalesStage: `Sales Stage`
   - See Figure 6.50

- Revenue, Factored Revenue, or Opportunity Count:

1. Multi-row card
2. Fields:
   - Fact: `Revenue` or `Factored Revenue` or `Opportunity Count`
   - See Figure 6.51

- AVG Revenue by Partner Driven and Opportunity Size:

1. Clustered bar chart
2. Fields:
   - Fact: `AVG Revenue`

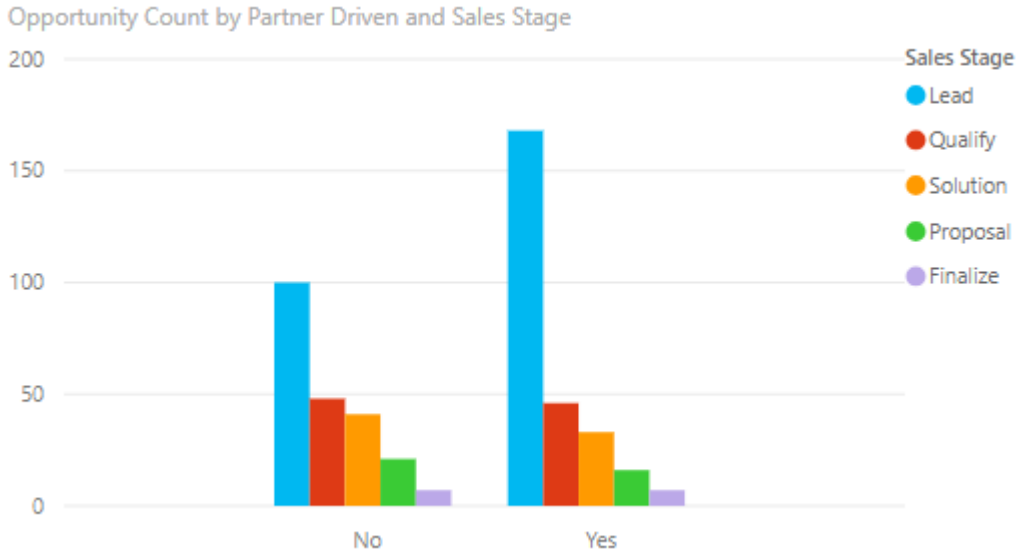


Figure 6.46:

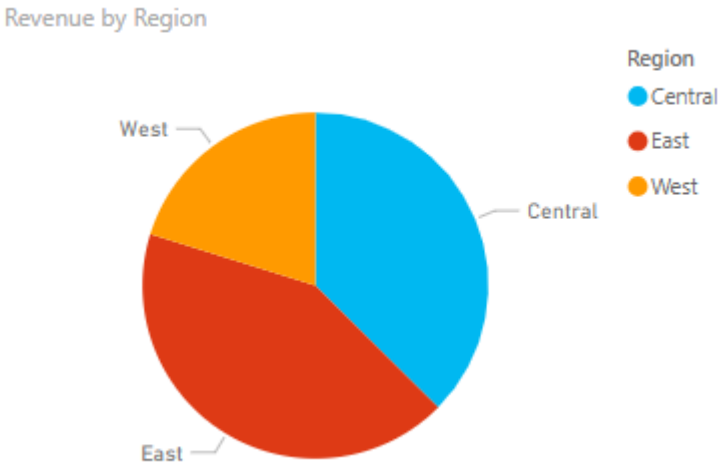


Figure 6.47:

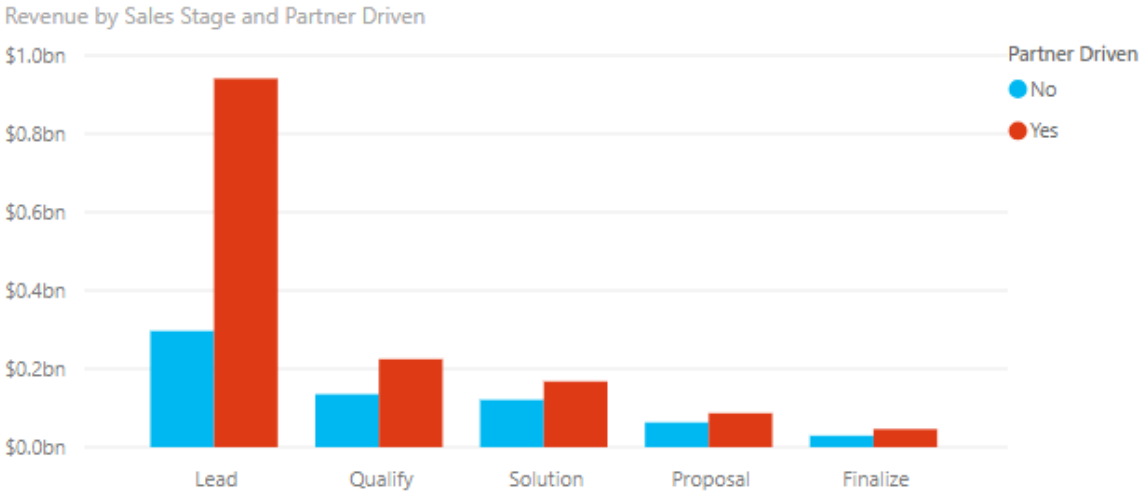


Figure 6.48:

Figure 6.49:

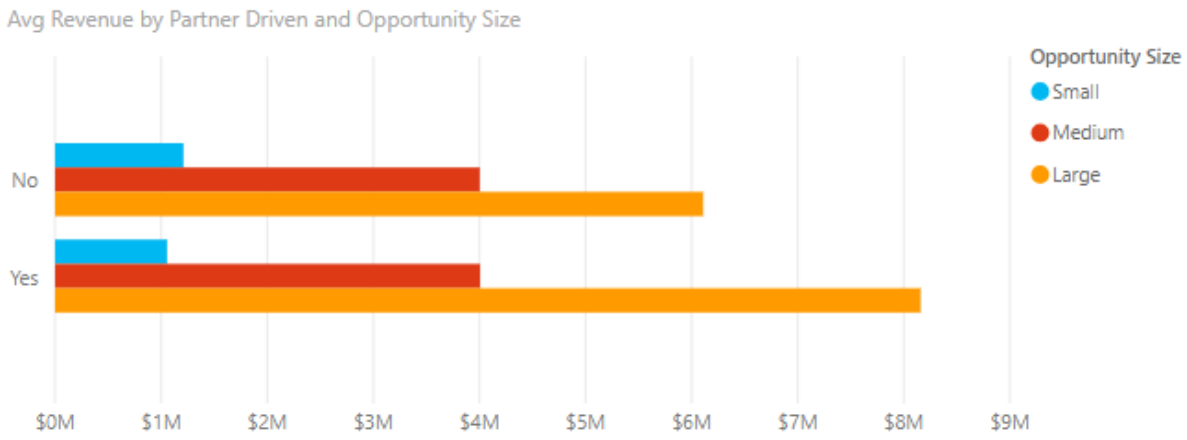


Figure 6.50:

Figure 6.51:

- Opportunity: `Opportunity Size`
- Partner: `Partner Driven`
- See Figure 6.52

Upcoming Opportunities by Month Dashboard

- Opportunity Count:

1. Multi-row card
2. Fields:
   - Fact: `Oppostunity Count`
   - See Figure 6.53

- Factored Revenue by Opportunity Size:

1. Clustered column chart
2. Fields:
   - Fact: `Factored Revenue`
   - Opportunity: `Opportunity Size`
   - See Figure 6.54

- AVG Revenue by Partner Driven and Sales Stage:

1. Clustered bar chart
2. Fields:
   - Fact: `AVG Revenue`
   - Partner: `Partner Driven`
   - SalesStage: `Sales Stage`
   - See Figure 6.55

- Opportunity Count by Month and Sales Stage:

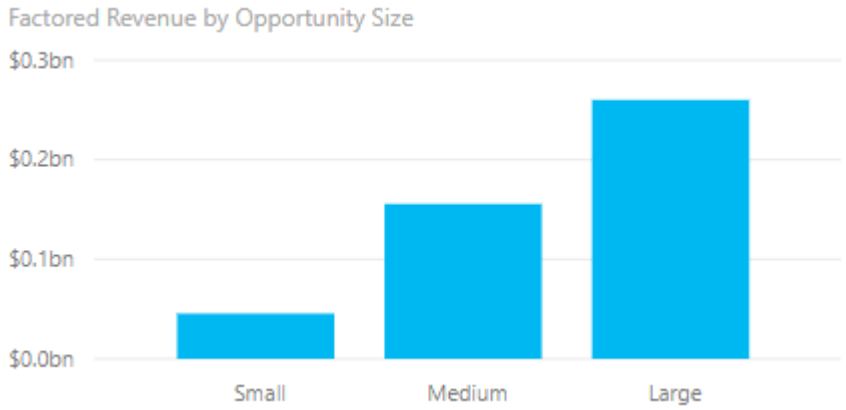


Figure 6.52:

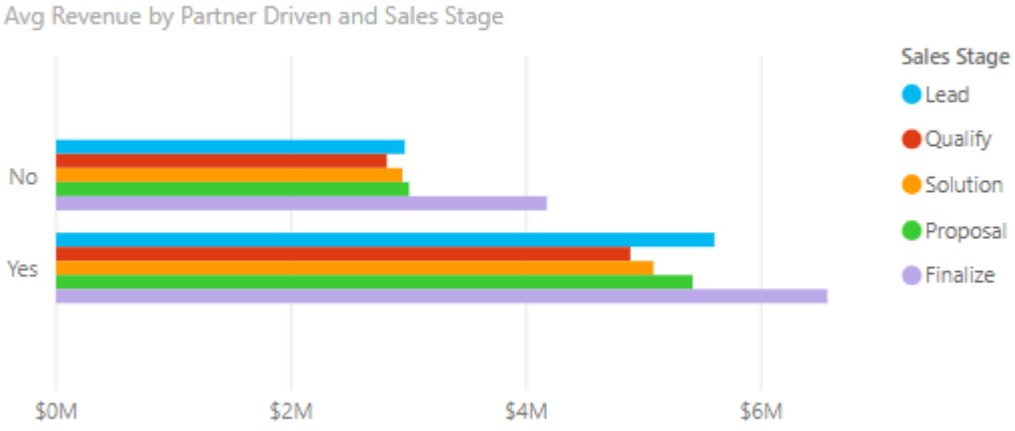


Figure 6.53:

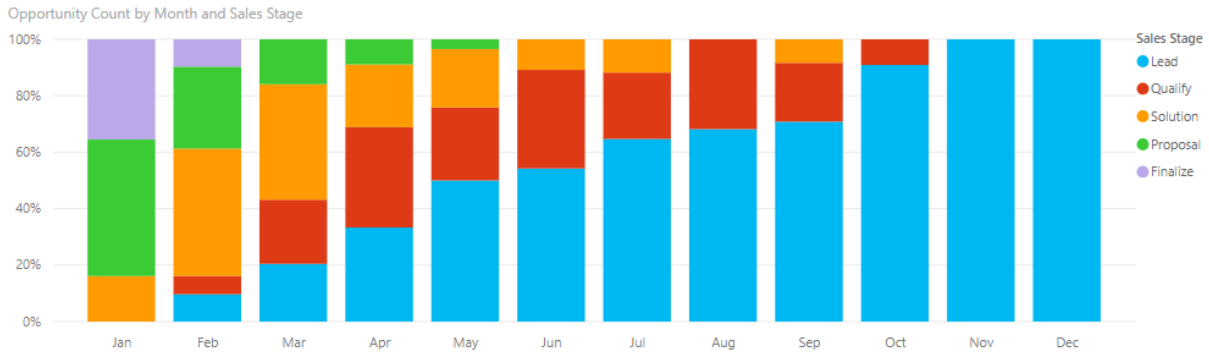


Figure 6.54:

1. 100 % Stacked column chart
2. Fields:
   - Fact: `Month`, `Oppostunity Count`
   - SalesStage: `Sales Stage`
   - See Figure 6.56

6.3.6 Sources

- `Opportunity Analysis sample for Power BI [76]`

6.4 Data Science HR Turnover

6.4.1 Introduction

- High employee retention rate (quit job) is big problem for companies.
- High employee retention rate (turnover or quit) costed a lot, fom Job postings, hiring processes, paperwork to new hire training
- For customer facing business such as E-commerce or consulting company, customers are often prefer to interact with familiar people. Errors and issues are more often likely if the company have new workers.

6.4.2 Litelature Review

The Top 5 reasons for job change:

1) Quality of Work
2) Performance Pressure
3) Friends
4) Location
5) Money

6.4.3 Data Description

Why are they best and most experienced employees leaving?. The dataset come from kaggle competition (https://www.kaggle.com/arvindbhatt/hrcsv).

Data Definition:

1. Satisfaction Level : Employee Satisfaction (can be interpreted as a %)

2. Last evaluation : Employee Evaluation (can be interpreted as a %)
3. Projects : Number of Projects (per year)
4. Average monthly hours : Average monthly hours
5. Time spent at company : Time spent at company
6. Accident : Whether they have had a work accident
7. Promotion Last 5 yrs : Whether they have had a promotion in the last 5 years
8. Department : Type of Job Position
9. Salary : Salary level (1= low, 2= medium, 3= high)
10. Left : Whether the employee has left (0= remains employed, 1= left)

6.4.4 Model Analysis

What factors increase job satisfaction?

In this case we used regression model (linear) to know the significant value of parameters.

Import dataset:

```
hr = read.csv("hr.csv")  # read csv file
```

Exploratory Data Analysis:

An Automated EDA with package `dlookr` and `DataExplorer`. I created a website about this EDA with github pages. More about github pages, see the site here: https://pages.github.com/

`dlookr`: diagnose(). See the result here: https://itsmecevi.github.io/eda-dlookr/

`DataExplorer`: create_report(). See the result here: https://itsmecevi.github.io/eda-dataexplorer/

Full Model:

Nine independent (9) variables in the dataset predicted the one (1) dependent variable (satisfaction level).

```
head(hr)
```

```
##   satisfaction_level last_evaluation number_project average_montly_hours
## 1               0.38            0.53              2                   157
## 2               0.80            0.86              5                   262
## 3               0.11            0.88              7                   272
## 4               0.72            0.87              5                   223
## 5               0.37            0.52              2                   159
## 6               0.41            0.50              2                   153
##   time_spend_company Work_accident left promotion_last_5years Department
## 1                  3             0    1                     0      sales
## 2                  6             0    1                     0      sales
## 3                  4             0    1                     0      sales
## 4                  5             0    1                     0      sales
## 5                  3             0    1                     0      sales
## 6                  3             0    1                     0      sales
##   salary
## 1    low
## 2 medium
## 3 medium
## 4    low
## 5    low
```

```
## 6    low
```

```
fullmodel <- lm(satisfaction_level ~ salary + average_montly_hours + number_project + time_spend_company

summary(fullmodel)
```

```
##
## Call:
## lm(formula = satisfaction_level ~ salary + average_montly_hours +
##     number_project + time_spend_company + promotion_last_5years +
##     last_evaluation + Work_accident + left, data = hr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64740 -0.13677 -0.01193  0.17004  0.52773
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.148e-01  1.159e-02  53.065  < 2e-16 ***
## salarylow               1.200e-02  6.961e-03   1.724   0.0847 .
## salarymedium            1.306e-02  6.956e-03   1.878   0.0604 .
## average_montly_hours    1.913e-04  4.127e-05   4.636 3.58e-06 ***
## number_project         -4.090e-02  1.691e-03 -24.183  < 2e-16 ***
## time_spend_company     -5.525e-03  1.295e-03  -4.267 2.00e-05 ***
## promotion_last_5years   9.285e-03  1.272e-02   0.730   0.4655
## last_evaluation         2.460e-01  1.167e-02  21.071  < 2e-16 ***
## Work_accident          -3.356e-05  5.238e-03  -0.006   0.9949
## left                   -2.241e-01  4.449e-03 -50.360  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2227 on 14989 degrees of freedom
## Multiple R-squared:  0.1982, Adjusted R-squared:  0.1977
## F-statistic: 411.6 on 9 and 14989 DF,  p-value: < 2.2e-16
```

Average monthly hours, number project, time spent at company, last evaluation, and whether or not the employee left have significant p-values.

R-squared and adjusted R-squared are both at about 0.19. This is not typically a good value, but is rather common in any data analysis of human behavior. For example, in psychology studies this R-squared level would not eliminate the model's validity, especially if p-values indicate significance

More abour summary function in linear model Regression:https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R

Revised Model:

```
revisedmodel<- lm(satisfaction_level ~ average_montly_hours + number_project + time_spend_company + las
summary(revisedmodel)
```

```
##
## Call:
## lm(formula = satisfaction_level ~ average_montly_hours + number_project +
##     time_spend_company + last_evaluation, data = hr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.61923 -0.19061  0.02274  0.19617  0.59000
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.152e-01  1.066e-02   57.70   <2e-16 ***
## average_montly_hours  5.183e-05  4.469e-05    1.16    0.246
## number_project       -3.894e-02  1.835e-03  -21.23   <2e-16 ***
## time_spend_company   -1.498e-02  1.383e-03  -10.83   <2e-16 ***
## last_evaluation       2.622e-01  1.266e-02   20.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2417 on 14994 degrees of freedom
## Multiple R-squared:  0.05522,    Adjusted R-squared:  0.05497
## F-statistic: 219.1 on 4 and 14994 DF,  p-value: < 2.2e-16
```

This model has even lower R-squared and adjusted R-squared values of about 0.05, which again can be attributed to the fact that human behavior is very difficult to predict. The variables remained significant, except for average monthly hours. Based on these p-values, we concluded the model and remove the insignificant variable for the next model.

6.4.5 R Code Analysis

DataSet Details:

```
dim(hr)
```

```
## [1] 14999    10
```

Describe DataSet:

Descriptive statistics (min, max, median etc) of each variable.

```
# > describe(hr)
#                       vars     n   mean    sd median trimmed   mad   min max  range  skew kurtosis   
# satisfaction_level       1 14999   0.61  0.25   0.64    0.63  0.28  0.09   1   0.91 -0.48    -0.67 0.0
# last_evaluation          2 14999   0.72  0.17   0.72    0.72  0.22  0.36   1   0.64 -0.03    -1.24 0.0
# number_project           3 14999   3.80  1.23   4.00    3.74  1.48  2.00   7   5.00  0.34    -0.50 0.0
# average_montly_hours     4 14999 201.05 49.94 200.00  200.64 65.23 96.00 310 214.00  0.05    -1.14 0.4
# time_spend_company       5 14999   3.50  1.46   3.00    3.28  1.48  2.00  10   8.00  1.85     4.77 0.0
# Work_accident            6 14999   0.14  0.35   0.00    0.06  0.00  0.00   1   1.00  2.02     2.08 0.0
# left                     7 14999   0.24  0.43   0.00    0.17  0.00  0.00   1   1.00  1.23    -0.49 0.0
# promotion_last_5years    8 14999   0.02  0.14   0.00    0.00  0.00  0.00   1   1.00  6.64    42.03 0.0
# Department*              9 14999   6.94  2.75   8.00    7.23  2.97  1.00  10   9.00 -0.79    -0.62 0.0
# salary*                 10 14999   2.35  0.63   2.00    2.41  1.48  1.00   3   2.00 -0.42    -0.67 0.0
```

Create a Contigency Table for each variable in dataset:

```
table_salary<-with(hr,table(salary))

table_salary
```

```
## salary
##   high    low medium
##   1237   7316   6446
```

```
table_satisfication<-with(hr,table(satisfaction_level))
table_satisfication
```

```
## satisfaction_level
## 0.09  0.1 0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19  0.2 0.21 0.22 0.23
##  195  358  335   30   54   73   76   79   72   63   74   69   67   60   54
## 0.24 0.25 0.26 0.27 0.28 0.29  0.3 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38
##   80   34   30   30   31   38   39   59   50   36   48   37  139  241  189
## 0.39  0.4 0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49  0.5 0.51 0.52 0.53
##  175  209  171  155  224  211  203   95   42  149  209  229  187  196  179
## 0.54 0.55 0.56 0.57 0.58 0.59  0.6 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68
##  185  179  187  210  182  219  193  208  188  209  187  199  228  177  162
## 0.69  0.7 0.71 0.72 0.73 0.74 0.75 0.76 0.77 0.78 0.79  0.8 0.81 0.82 0.83
##  209  205  171  230  246  257  226  234  252  241  217  222  220  241  234
## 0.84 0.85 0.86 0.87 0.88 0.89  0.9 0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98
##  247  207  200  225  187  237  220  224  198  169  167  181  203  176  183
## 0.99    1
##  172  111
```

```
table_lastevaluation<-with(hr,table(last_evaluation))
table_lastevaluation
```

```
## last_evaluation
## 0.36 0.37 0.38 0.39  0.4 0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49  0.5
##   22   55   50   52   57   59   56   50   44  115  211  173  292  332  353
## 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59  0.6 0.61 0.62 0.63 0.64 0.65
##  345  309  324  350  358  322  333  225  255  221  234  233  236  235  201
## 0.66 0.67 0.68 0.69  0.7 0.71 0.72 0.73 0.74 0.75 0.76 0.77 0.78 0.79  0.8
##  222  245  222  193  213  196  211  223  260  238  216  263  214  241  251
## 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89  0.9 0.91 0.92 0.93 0.94 0.95
##  255  237  269  294  316  273  326  235  296  313  287  269  269  263  258
## 0.96 0.97 0.98 0.99    1
##  249  276  263  258  283
```

```
table_numberproject<-with(hr,table(number_project))
table_numberproject
```

```
## number_project
##    2    3    4    5    6    7
## 2388 4055 4365 2761 1174  256
```

```
table_avgmontlyhours<-with(hr,table(average_montly_hours))
table_avgmontlyhours
```

```
## average_montly_hours
##  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113
##   6  14  23  11  19  16  17  17  28  17  19  10  18  18  12  26  10  29
## 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131
##  15  14  10  18  12  10  10  24  11  20  13  19  25  72  65  63  59  69
## 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149
## 100  87 114 153 104 122  88 120 129 115 112 127 102 134 110 118 123 148
## 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167
## 108 147 112 122 121 125 153 126 124 121 136  87  96  73  78  78  73  94
## 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185
##  92  86  76  83  70  96  78  76  81  81  85  73  88  78  75  84  80  93
## 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203
##  76  68  73  85  75  80  96  67  71  67  79  70  86  79  58  86  80  72
## 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221
##  68  73  83  71  72  72  72  79  72  71  78  68  76  87  79  85  64  81
```

```
## 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239
##  84  93 112  95  93  77  76  93  59  77  97 102  74  76  83  90 108  96
## 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257
##  93  85  98 112  98 124 102 108  86  93 100  98  86 101 113 115  87 126
## 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275
## 110  98 124 102  86 110 111  91 105  88  93 102  93 104  86  88  94  82
## 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293
##  30  21  35  32  29  34  36  25  24  33  50  30   6  19  15  17  15  13
## 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310
##  16  12  21   7  13   6  11  24   8   6  17  18  18  14  20  16  18
```

```
table_timespend<-with(hr,table(time_spend_company))
table_timespend
```

```
## time_spend_company
##    2    3    4    5    6    7    8   10
## 3244 6443 2557 1473  718  188  162  214
```

```
table_workaccident<-with(hr,table(Work_accident))
table_workaccident
```

```
## Work_accident
##     0     1
## 12830  2169
```

```
table_left<-with(hr,table(left))
table_left
```

```
## left
##     0     1
## 11428  3571
```

```
table_promotion<-with(hr,table(promotion_last_5years))
table_promotion
```

```
## promotion_last_5years
##     0     1
## 14680   319
```

```
table_sales<-with(hr,table(Department))
table_sales
```

```
## Department
##  accounting          hr          IT  management   marketing product_mng
##         767         739        1227         630         858         902
##       RandD       sales     support   technical
##         787        4140        2229        2720
```

```
table_salary<-with(hr,table(salary))
table_salary
```

```
## salary
##   high    low medium
##   1237   7316   6446
```

```
table_promotion_salary<-xtabs(~promotion_last_5years+salary,data=hr)
table_promotion_salary
```

```
##                       salary
```

```
## promotion_last_5years high  low medium
##                     0 1165 7250   6265
##                     1   72   66    181
```

```
table_project_timespend<-xtabs(~number_project+time_spend_company,data=hr)
table_project_timespend
```

```
##                time_spend_company
## number_project    2    3    4    5    6    7    8   10
##              2  224 1854  136   83   53   16   12   10
##              3 1255 1782  530  135  139   58   62   94
##              4 1144 1798  577  445  215   64   46   76
##              5  554  866  431  592  224   38   34   22
##              6   66  136  673  180   87   12    8   12
##              7    1    7  210   38    0    0    0    0
```

Boxplot Creation:

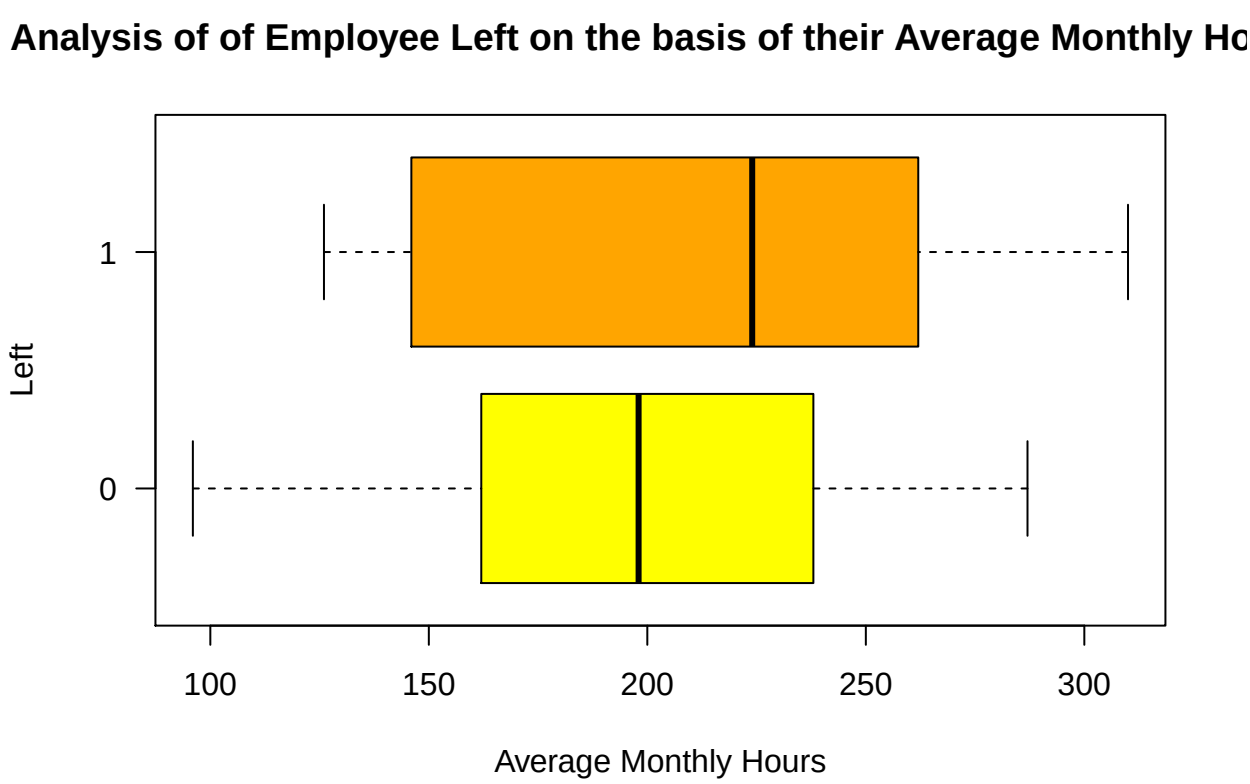
```
boxplot(satisfaction_level ~salary,data=hr, horizontal=TRUE,
          ylab="Salary Level", xlab="Satisfaction level", las=1,
          main="Analysis of Salary of Employee on the basis of their satisfaction level",
          col=c("red","blue","green")
          )
```


```
boxplot(satisfaction_level ~left, data=hr, horizontal=TRUE,
          ylab="Left", xlab="Satisfaction level", las=1,
          main="Analysis of of Employee Left on the basis of their satisfaction level",
          col=c("Yellow","Orange")
```

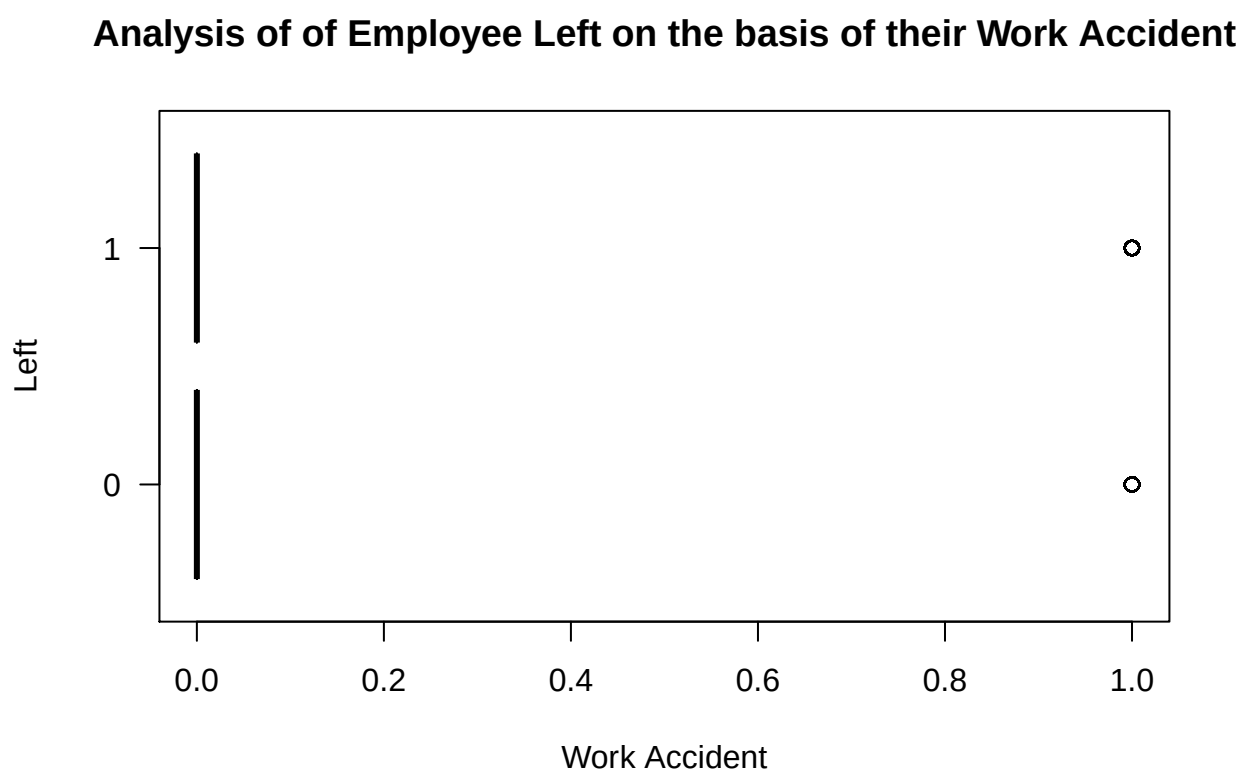
```
          )
```


```
boxplot(number_project~left,data=hr, horizontal=TRUE,
          ylab="Left", xlab="No of Projects", las=1,
          main="Analysis of of Employee Left on the basis of their Number of Projects",
          col=c("Red","Magenta")
          )
```


```
boxplot(average_montly_hours ~left, data=hr,horizontal=TRUE,
          ylab="Left", xlab="Average Monthly Hours", las=1,
          main="Analysis of of Employee Left on the basis of their Average Monthly Hours",
          col=c("Yellow","Orange")
          )
```



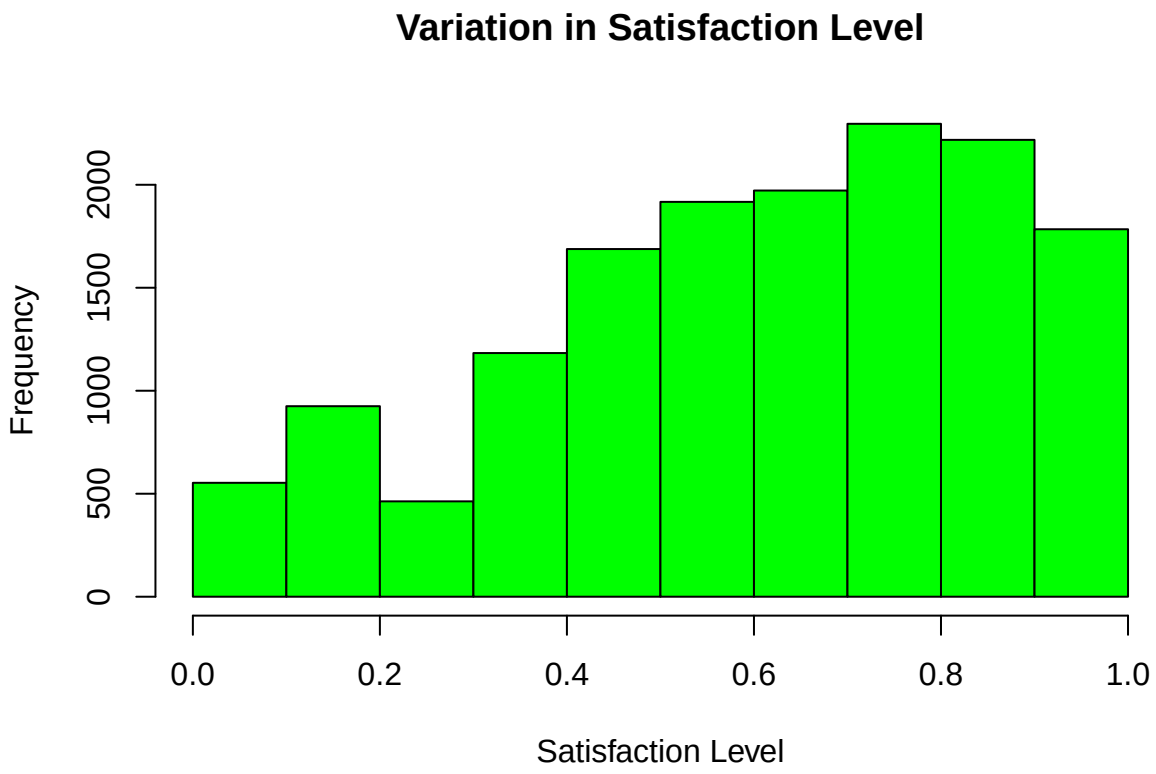

```
boxplot(Work_accident~left,data=hr, horizontal=TRUE,
          ylab="Left", xlab="Work Accident", las=1,
          main="Analysis of of Employee Left on the basis of their Work Accident",
          col=c("Yellow","Orange")
          )
```



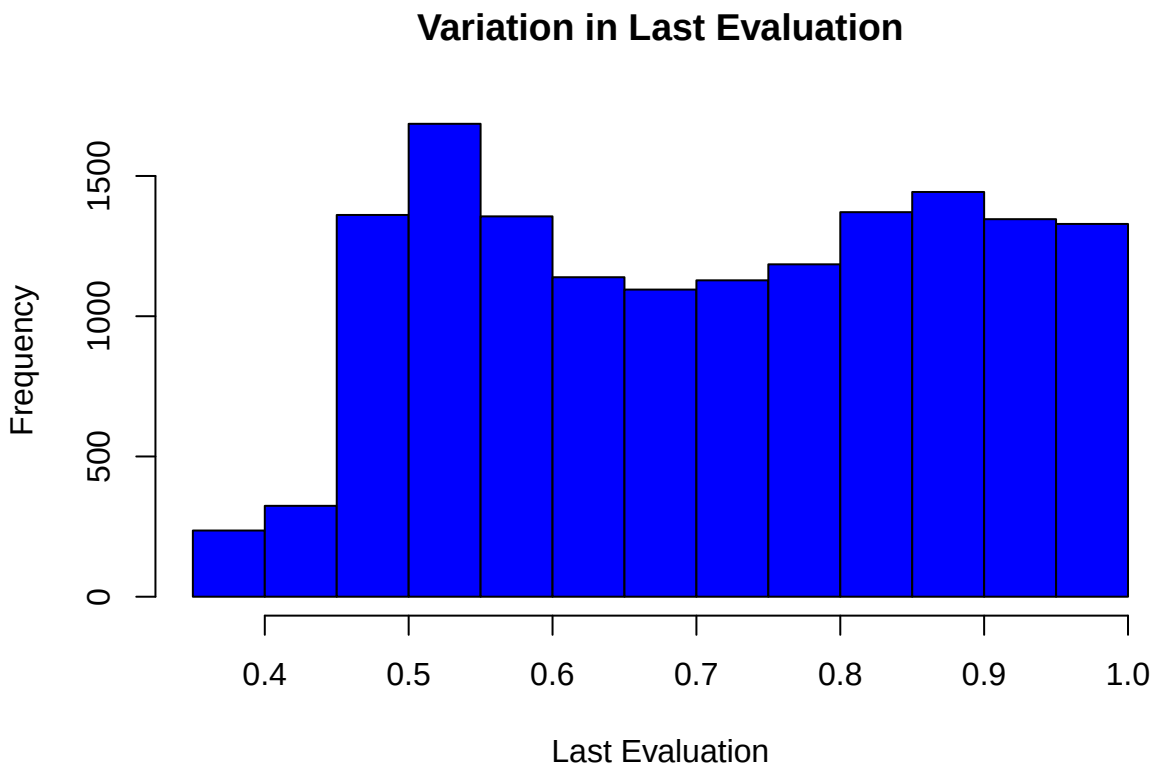

```
boxplot(last_evaluation ~left,data=hr, horizontal=TRUE,
          ylab="Left", xlab="Last Evaluation", las=1,
          main="Analysis of of Employee Left on the basis of their Last Evaluation",
          col=c("Yellow","Orange")
          )
```

Histogram for the variables:

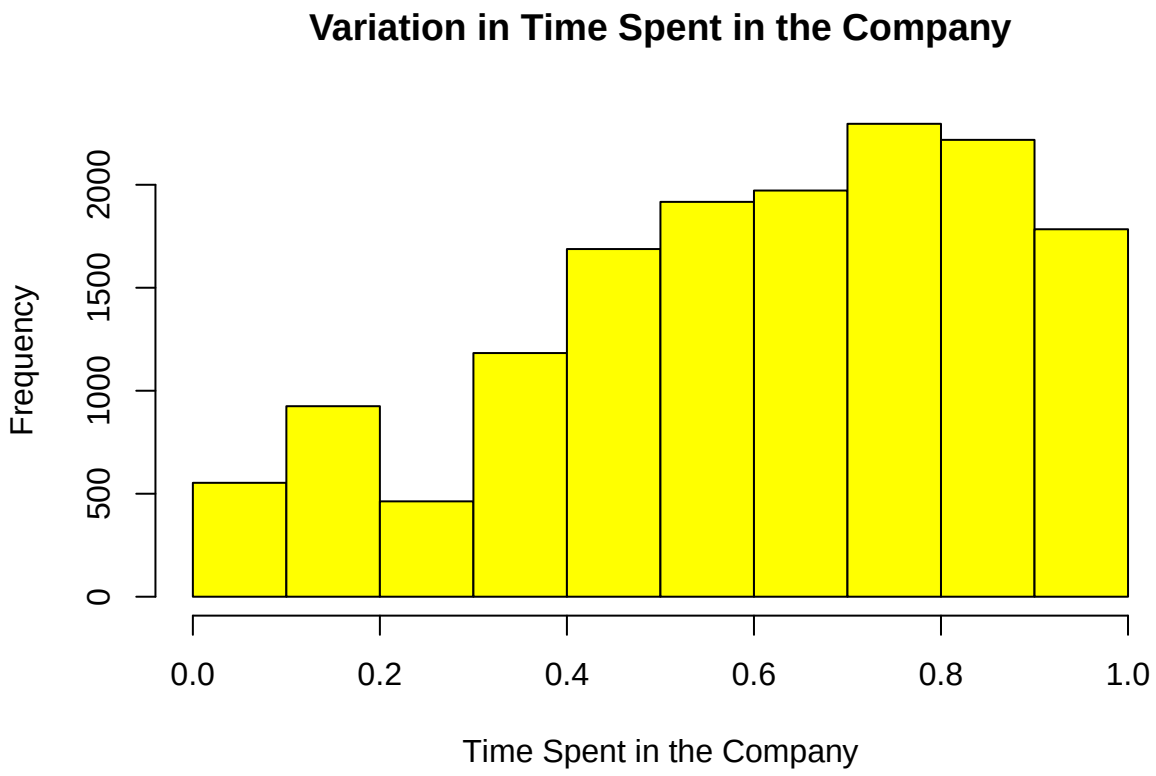
```
hist(hr$satisfaction_level, main=" Variation in Satisfaction Level ", xlab="Satisfaction Level",breaks=
```



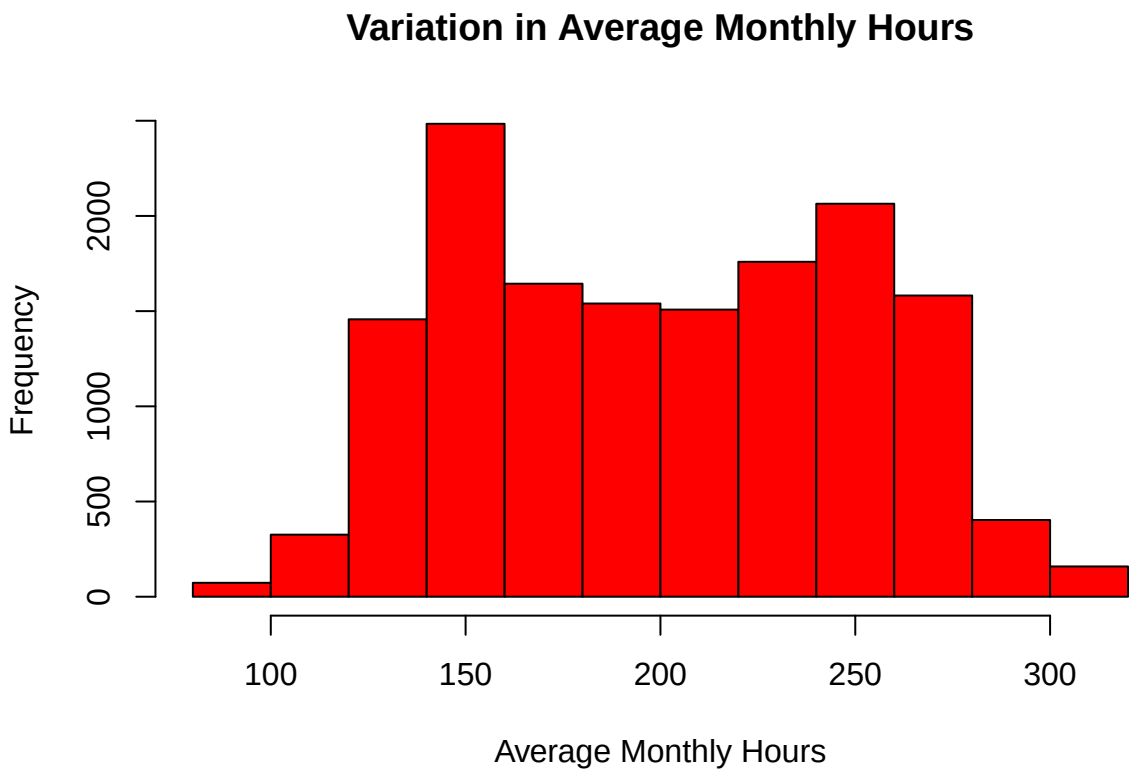

```
hist(hr$last_evaluation, main=" Variation in Last Evaluation ", xlab="Last Evaluation",breaks=10,ylab="
```



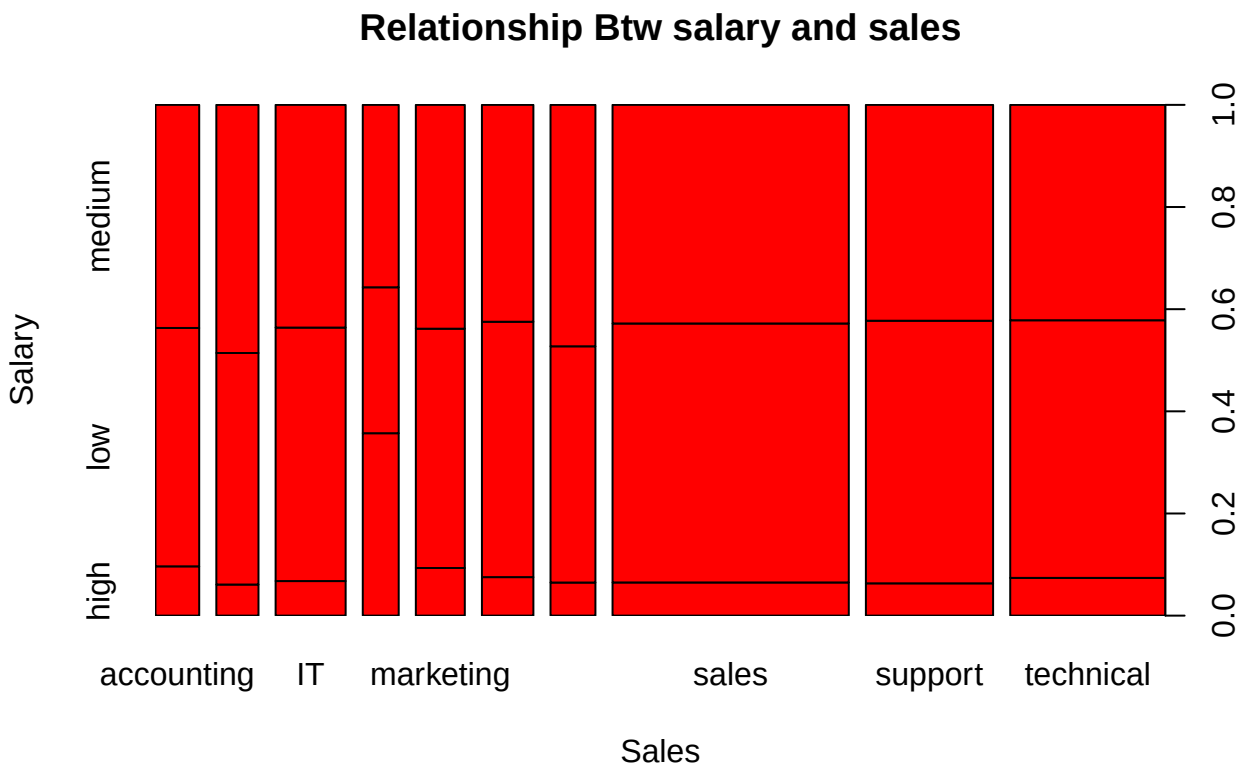

```
hist(hr$satisfaction_level, main=" Variation in Time Spent in the Company ", xlab="Time Spent in the Co
```



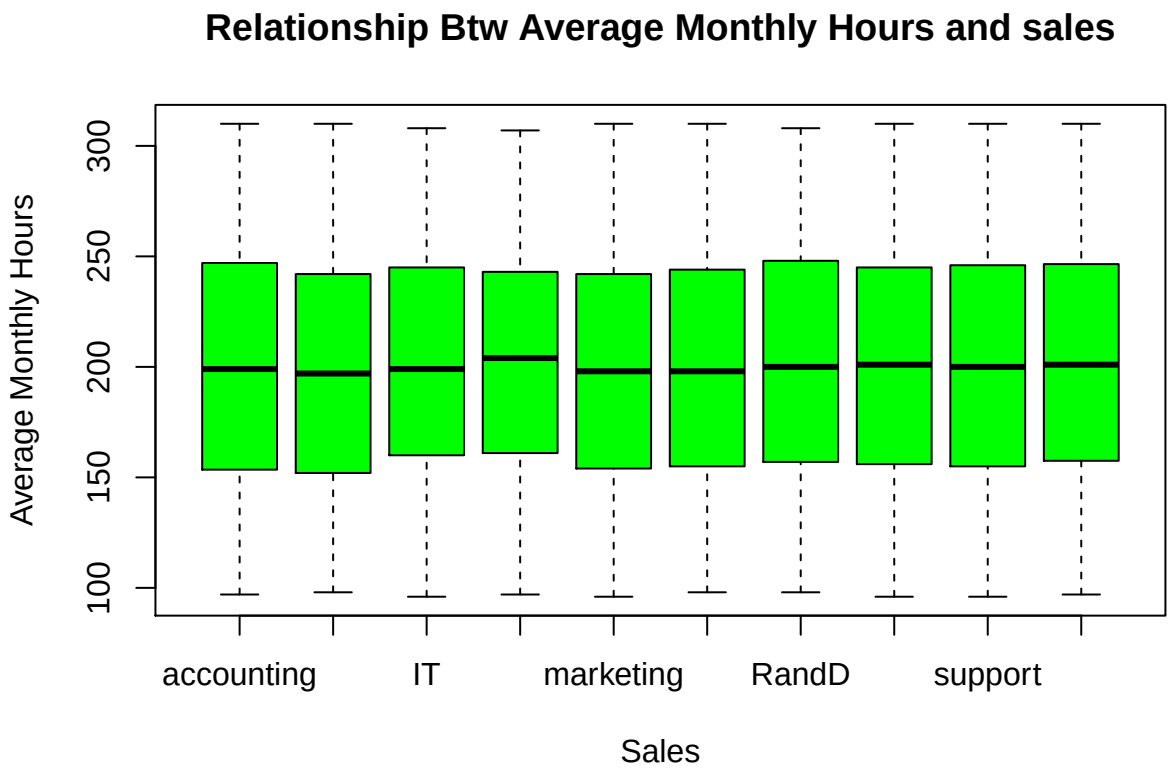

```
hist(hr$average_montly_hours, main=" Variation in Average Monthly Hours ", xlab="Average Monthly Hours"
```




```
plot(y=hr$salary, x=hr$Department,
     col="red",
     main="Relationship Btw salary and sales",
     ylab="Salary", xlab="Sales")
```




```
plot(y=hr$average_montly_hours, x=hr$Department,
     col="green",
     main="Relationship Btw Average Monthly Hours and sales",
     ylab="Average Monthly Hours", xlab="Sales")
```



Visualize the correlation matrix:

We need `corrplot` package.

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
#correlationMatrix <- cor(hr[,c(1:8)])
#corrplot(correlationMatrix, method="circle")
```

```
cor(hr[ ,c(1,2,3,4,5,6,7,8)])
```

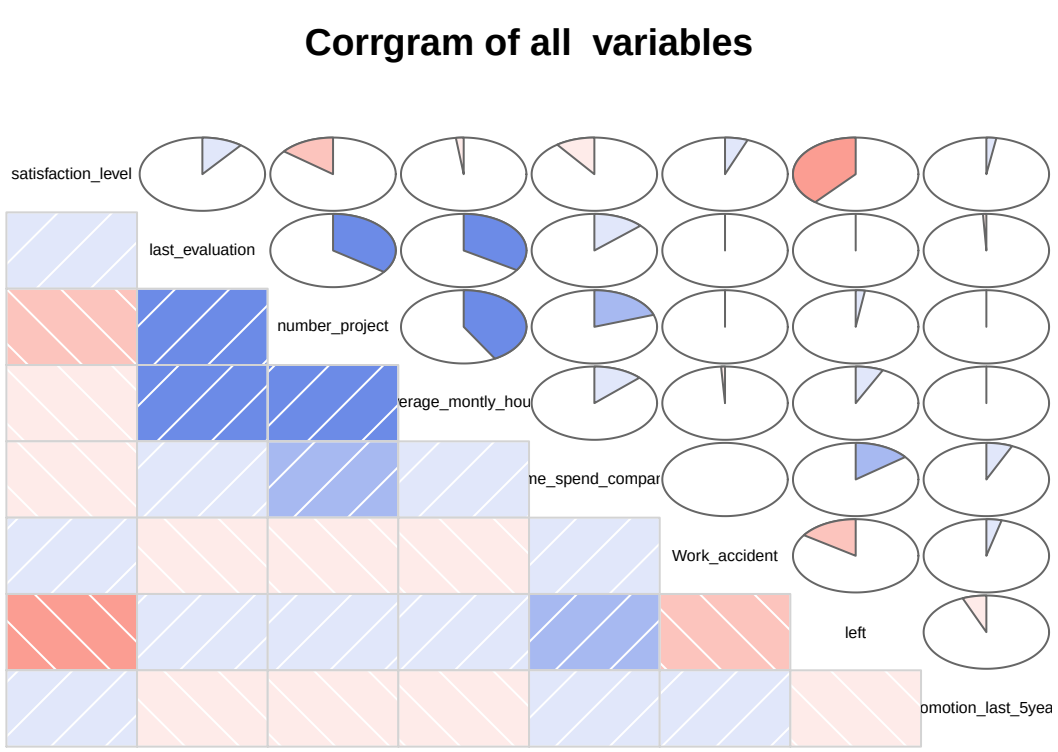
```
##                        satisfaction_level last_evaluation number_project
## satisfaction_level             1.00000000     0.105021214   -0.142969586
## last_evaluation                0.10502121     1.000000000    0.349332589
## number_project                -0.14296959     0.349332589    1.000000000
## average_montly_hours          -0.02004811     0.339741800    0.417210634
## time_spend_company            -0.10086607     0.131590722    0.196785891
## Work_accident                  0.05869724    -0.007104289   -0.004740548
## left                          -0.38837498     0.006567120    0.023787185
## promotion_last_5years          0.02560519    -0.008683768   -0.006063958
##                        average_montly_hours time_spend_company
## satisfaction_level           -0.020048113       -0.100866073
## last_evaluation               0.339741800        0.131590722
## number_project                0.417210634        0.196785891
## average_montly_hours          1.000000000        0.127754910
## time_spend_company            0.127754910        1.000000000
```

```
## Work_accident                -0.010142888        0.002120418
## left                          0.071287179        0.144822175
## promotion_last_5years        -0.003544414        0.067432925
##                         Work_accident        left promotion_last_5years
## satisfaction_level       0.058697241 -0.38837498           0.025605186
## last_evaluation         -0.007104289  0.00656712          -0.008683768
## number_project          -0.004740548  0.02378719          -0.006063958
## average_montly_hours    -0.010142888  0.07128718          -0.003544414
## time_spend_company       0.002120418  0.14482217           0.067432925
## Work_accident            1.000000000 -0.15462163           0.039245435
## left                    -0.154621634  1.00000000          -0.061788107
## promotion_last_5years    0.039245435 -0.06178811           1.000000000
```

We need corrgram package.

```
library(corrgram)
```

```
corrgram(hr, lower.panel = panel.shade, upper.panel = panel.pie, text.panel = panel.txt, main = "Corrgra
```



Run a suitable test to check your hypothesis for your suitable assumptions:

```
cor.test(hr$left,hr$satisfaction_level)
```

```
##
##  Pearson's product-moment correlation
##
## data:  hr$left and hr$satisfaction_level
## t = -51.613, df = 14997, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4018809 -0.3747001
## sample estimates:
##       cor
## -0.388375
```

```
cor.test(hr$left,hr$last_evaluation)
```

```
##
##  Pearson's product-moment correlation
##
## data:  hr$left and hr$last_evaluation
## t = 0.80424, df = 14997, p-value = 0.4213
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.009437678  0.022568555
## sample estimates:
##        cor
## 0.00656712
```

```
cor.test(hr$left,hr$number_project)
```

```
##
##  Pearson's product-moment correlation
##
## data:  hr$left and hr$number_project
## t = 2.9139, df = 14997, p-value = 0.003575
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.007786343 0.039775850
## sample estimates:
##        cor
## 0.02378719
```

T- Test Hypothesis:

```
t.test(hr$satisfaction_level~hr$left)
```

```
##
##  Welch Two Sample t-test
##
## data:  hr$satisfaction_level by hr$left
## t = 46.636, df = 5167, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2171815 0.2362417
## sample estimates:
## mean in group 0 mean in group 1
##       0.6668096       0.4400980
```

```
t.test(hr$time_spend_company~hr$left)
```

```
##
##  Welch Two Sample t-test
##
## data:  hr$time_spend_company by hr$left
```

```
## t = -22.631, df = 9625.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5394767 -0.4534706
## sample estimates:
## mean in group 0 mean in group 1
##        3.380032        3.876505
```

```
t.test(hr$average_montly_hours~hr$left)
```

```
##
##  Welch Two Sample t-test
##
## data:  hr$average_montly_hours by hr$left
## t = -7.5323, df = 4875.1, p-value = 5.907e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.534631  -6.183384
## sample estimates:
## mean in group 0 mean in group 1
##        199.0602        207.4192
```

```
t.test(hr$last_evaluation~hr$left)
```

```
##
##  Welch Two Sample t-test
##
## data:  hr$last_evaluation by hr$left
## t = -0.72534, df = 5154.9, p-value = 0.4683
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.009772224  0.004493874
## sample estimates:
## mean in group 0 mean in group 1
##       0.7154734       0.7181126
```

6.4.6 Result

1. Why are our best and most experienced employees leaving prematurely?

- High salaried employees show a different pattern for leaving the company
- The parameters to quit out job are
  - Satisfaction level < 0.5
  - Average monthly hours > 200
  - After spending average 4 years of time in the company

2. Which employee will leave next?

Emplpoyee who has low salary, with satisfaction level < 0.5 and is putting in average monthly hours > 200. The probability of the employee leaving is 70% (3 parameters from total 10 parameters)

3. Interesting Insights:

- Number of projects, time spent at company, and last evaluation are significant predictors of job satisfaction
- The well-balanced worker who has recently been promoted is the happiest
- Employees who are over-worked or under-work are relatively dissatisfied
- Employees at the lowest salary level (level 1) are actually the most satisfied

- The happiest employees work 3 or 4 projects each year
- Employeea with the most extreme hours are typically very disatisfied
- Employees with poor or excellent last evaluations are the most satisfied
- Employees who have left the company were typically on the higher end of average monthly hours
- Employees who left worked are under 3 projects, recieved poor or excellent performance rating, were not promoted in the last 5 years, and most did not have a work accident
- We see that the last evaluations of employees who left have a bimodal distribution with great quantities at each extreme
- In terms of time spent at the company, there is a huge spike in satisfaction levels for employees who have remained with the company for 2.5 years or more
- Satisfaction levels decrease past the 2.5 year satisfaction peak
- Loyal employees who have remained employed with the company for more than 6 years tend to be happier in their positions.

6.4.7 Sources

- `Human Resource Analytics Dataset [63]`
- `HUMAN RESOURCE ANALYSIS [64]`
- `HR Analytics- exploration and modelling with R [65]`
- `What Is the Meaning of Attrition Used in HR? [66]`
- `Top 5 Reasons for Employee Attrition & How to deal with it [67]`

6.5 Data Science Sales Prediction

6.5.1 Background and objective of the analysis

Note: This is just an example of use case (fake situation).

Some of the research says, that there is a strong positive correlation about train ticketing sales and homes sales of the region. Deutsche Bahn got the project to build Railways infrastructures from the USA. To know which line will build up and how many traffic in some region, Deutsche Bahn tried to find the home sales of that region. The result of this analytics is just little experiment and need further inspections. The data is available at the following URL: https://bit.ly/2U9QyPo

6.5.2 Reading the data

Importing data:

```
#First, please setwd() to our file directory
sales<-read.csv(file="sales.csv", header=TRUE,sep = ",")
head(sales, 5) #nur 5 Zeilen hat so gezeigt
```

```
##     Month
## 1 1973-01
## 2 1973-02
## 3 1973-03
## 4 1973-04
## 5 1973-05
##   Monthly.sales.of.new.one.family.houses.sold.in.th.e.USA.since.1973
## 1                                                                 55
## 2                                                                 60
## 3                                                                 68
## 4                                                                 63
## 5                                                                 65
```

```
dim(sales) #dimension der daten
```

```
## [1] 277   2
```

```
class(sales) #class der daten
```

```
## [1] "data.frame"
```

```
#Rename the column name
name<-c("month","sales")
names(sales)<-name
head(sales,5)
```

```
##     month sales
## 1 1973-01    55
## 2 1973-02    60
## 3 1973-03    68
## 4 1973-04    63
## 5 1973-05    65
```

So the data has 275 rows (277-2) and 2 columns (date and number of apartments) as "data.frame". It is CSV table formatting. Note! the data is monthly data from Jan 1973 - Nov 1995.

Create ts (time series) object:

The data is not `ts (time series) object` but `dataframe`. We have to convert the data to ts object. There are many methods to convert the data to `ts (time series) object`. One of them is `fpp`(Data for "Forecasting principles and practice") packages.

** Create times series object with fpp packagen: **

```
#install.packages("fpp")
#library(fpp)
sales_ts<-ts(sales$sales,frequency=12,start=c(1973))
sales_ts
```

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1973  55  60  68  63  65  61  54  52  46  42  37  30
## 1974  37  44  55  53  58  50  48  45  41  34  30  24
## 1975  29  34  44  54  57  51  51  53  46  46  46  39
## 1976  41  53  55  62  55  56  57  59  58  55  49  47
## 1977  57  68  84  81  78  74  64  74  71  63  55  51
## 1978  57  63  75  85  80  77  68  72  68  70  53  50
## 1979  53  58  73  72  68  63  64  68  60  54  41  35
## 1980  43  44  44  36  44  50  55  61  50  46  39  33
## 1981  37  40  49  44  45  38  36  34  28  29  27  29
## 1982  28  29  36  32  36  34  31  36  39  40  39  33
## 1983  44  46  57  59  64  59  51  50  48  51  45  48
## 1984  52  58  63  61  59  58  52  48  53  55  42  38
## 1985  48  55  67  60  65  65  63  61  54  52  51  47
## 1986  55  59  89  84  75  66  57  52  60  54  48  49
## 1987  53  59  73  72  62  58  55  56  52  52  43  37
## 1988  43  55  68  68  64  65  57  59  54  57  43  42
## 1989  52  51  58  60  61  58  62  61  49  51  47  40
## 1990  45  50  58  52  50  50  46  46  38  37  34  29
## 1991  30  40  46  46  47  47  43  46  37  41  39  36
## 1992  48  55  56  53  52  53  52  56  51  48  42  42
## 1993  44  50  60  66  58  59  55  57  57  56  53  51
```

```
## 1994  45  58  74  65  65  55  52  59  54  57  45  40
## 1995  47  47  60  58  63  64  64  63  55  54  44  NA
## 1996  NA
```

There are NA values, so we clean up right before the analysis. Because NA values destroys the aggregation. We need data without NA:

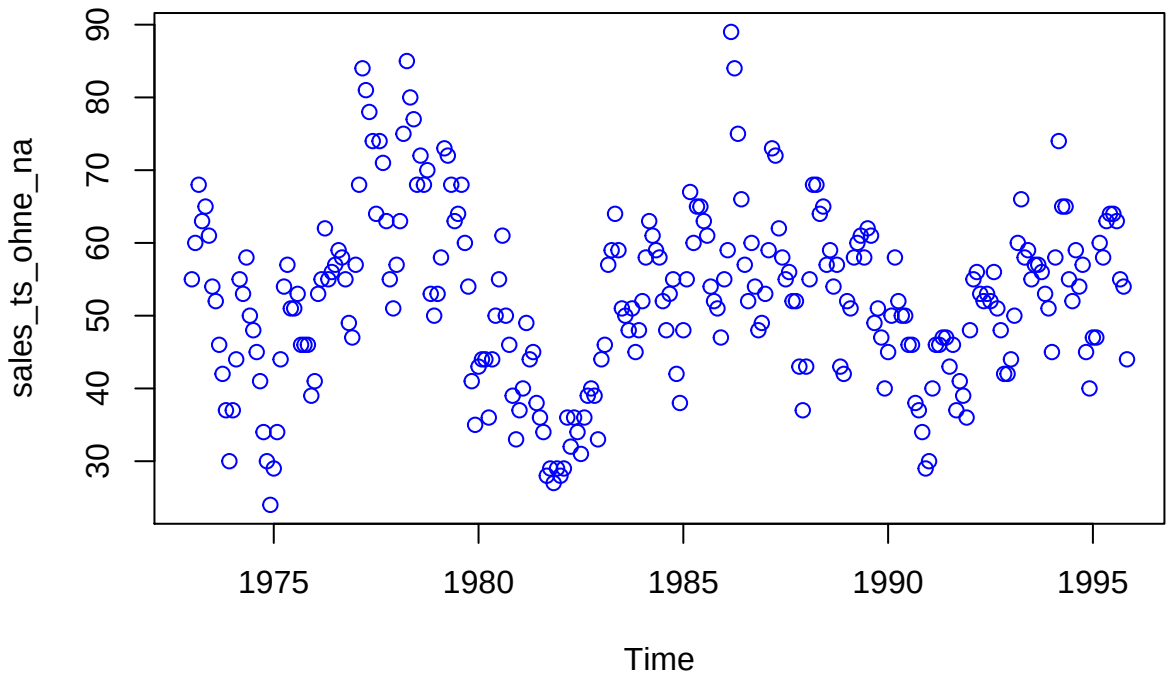
```
sales_ts_ohne_na<-na.remove(sales_ts) #Note! na.remove is in fpp package class(sales_ts_ohne_na)
```

```
sales_ts_ohne_na<-na.omit(sales_ts)
class(sales_ts_ohne_na)
```

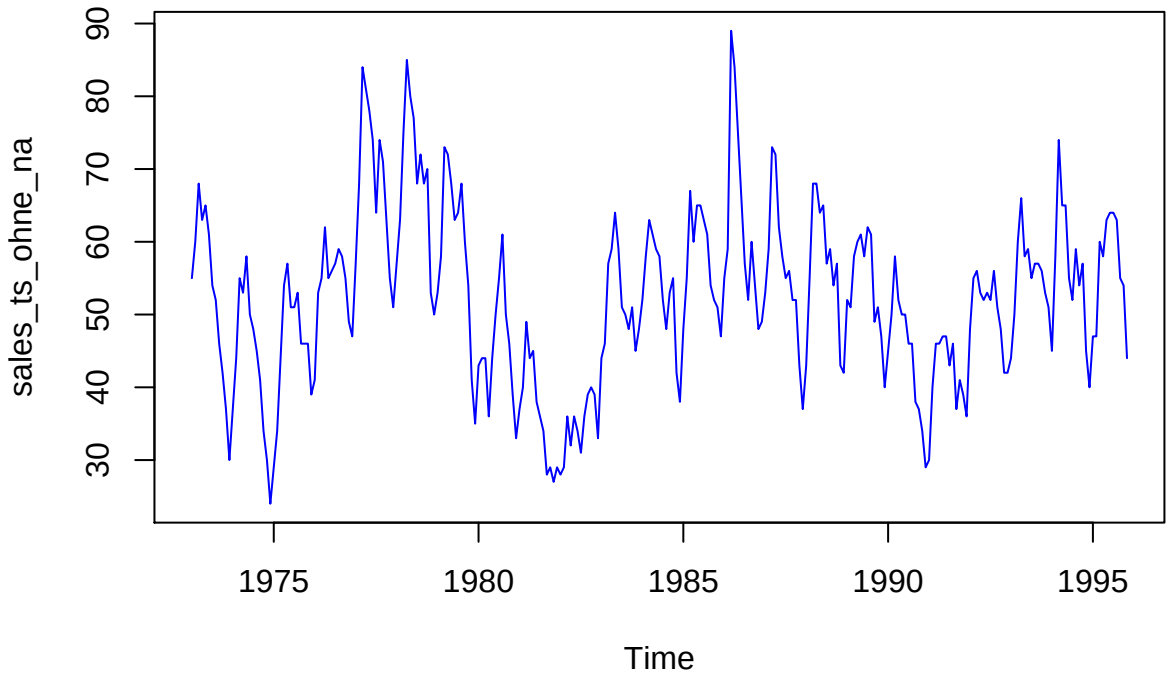
```
## [1] "ts"
```

Plot the data:

```
plot.ts(sales_ts_ohne_na, type="p", col="blue", lty="dashed") #scatterplot
```




```
ts.plot(sales_ts_ohne_na,col="blue") #Line graph
```



Add important statistical key figures to graph:

```
plot(sales_ts_ohne_na,col="blue")

#mean
mw=mean(sales_ts_ohne_na)
mw
```

```
## [1] 52.28727
```

```
#mean in graph
x=c(1973:1995,4)
#length(x)
y=rep(mw,24)
#length(y)
lines(x,y,col=2)

#Variance
var(sales_ts_ohne_na)
```

```
## [1] 142.534
```

```
#standard deviation (std)
std=sd(sales_ts_ohne_na)
std
```

```
## [1] 11.93876
```

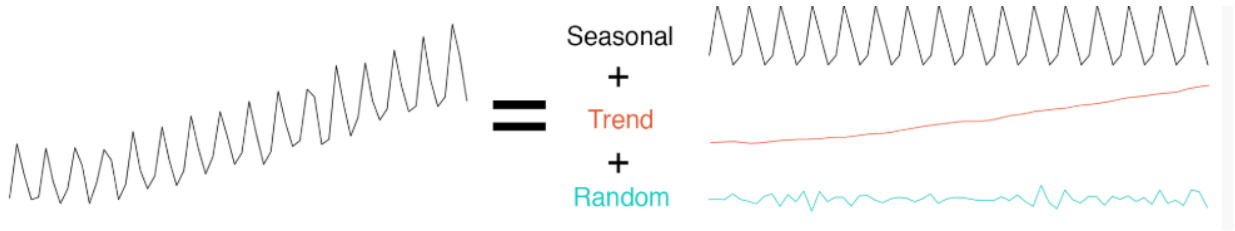
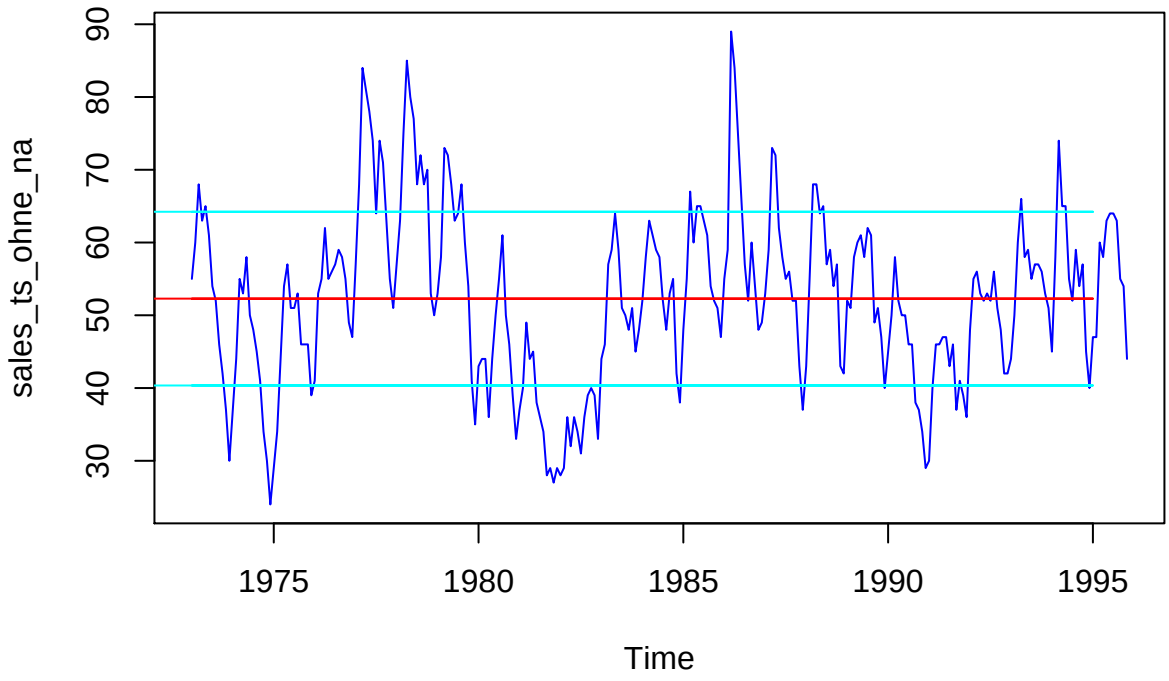


Figure 6.55:

```
# std
z1=rep(std+mw,24)
z2=rep(-std+mw,24)
#length(y)
lines(x,z1,col=5)
lines(x,z2,col=5)
```



6.5.3 Time series analysis

What are the components of a time series?

See Figure 5.67

Goal:

We want to work with stationary processes as far as possible, and even better with linear processes. But the process with

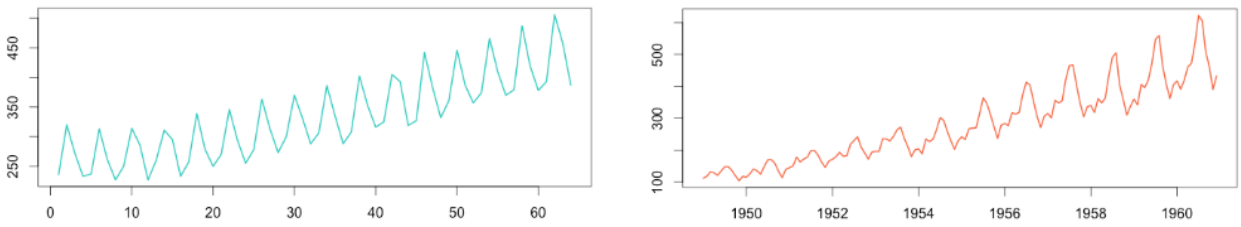


Figure 6.56:

- Trend
- Seasonal

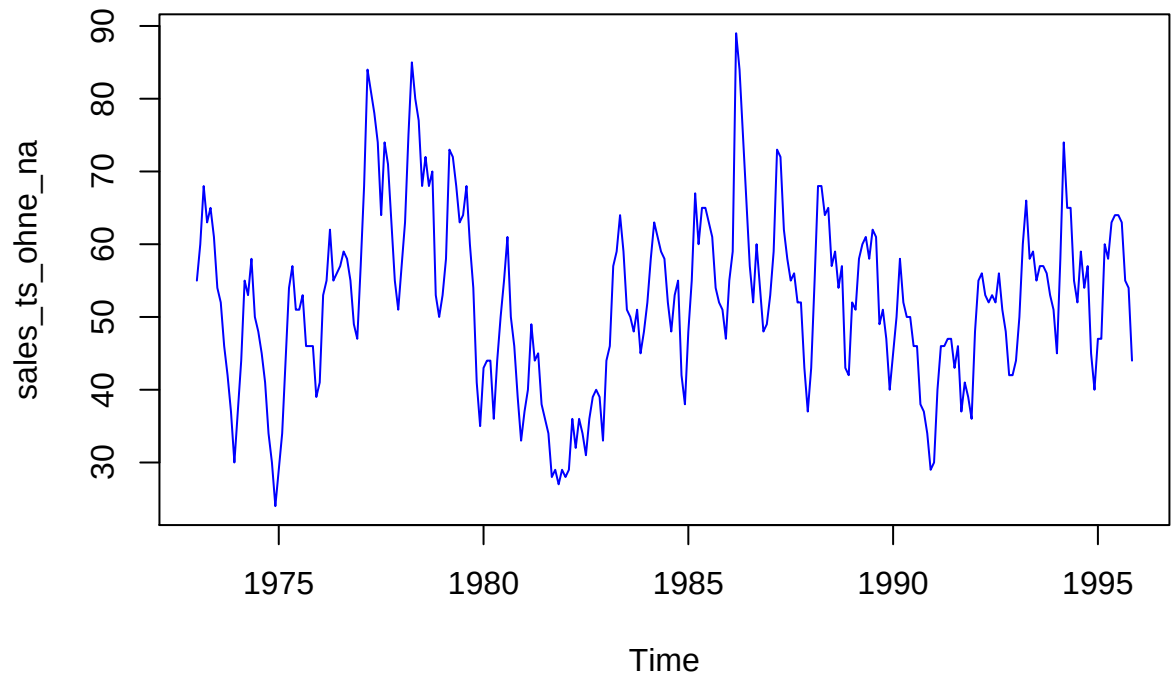
are not stationary. Therefore, we try to make them stationary. Trend and seasonal adjustment or decomposition.

Additive or Multiplicative Time Series?

In order to achieve a successful 'trend adjustment and seasonal adjustment', it is important to choose between the additive and the multiplicative model, which requires an analysis of the time series. See Figure 5.68 (Left: Additiv, Right: Multiplicative)

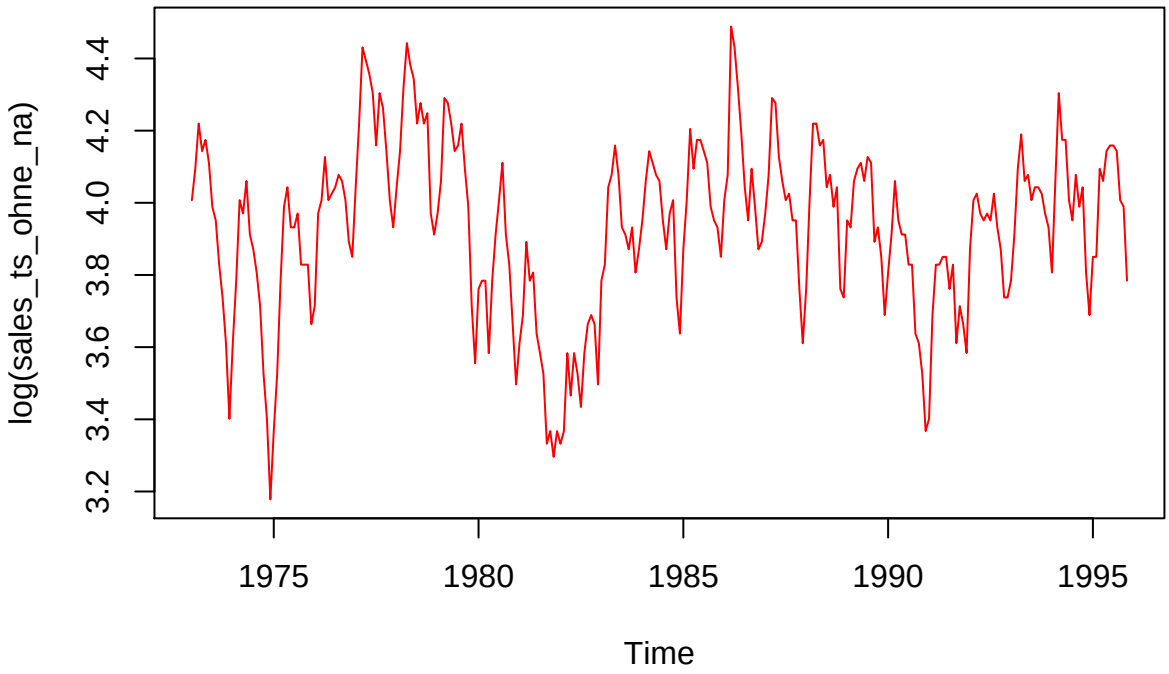
- Left: Additive: Time series = Seasonal + Trend + Random
- Right: Multiplicative: Time series = Trend * Seasonal * Random

```
plot.ts(sales_ts_ohne_na,col="blue") #without log function
```



To make symmetric, we insert log function.

```
plot.ts(log(sales_ts_ohne_na),col="red") #with log function
```



With and without log function ist the same graphs. we take the data without log function.

Step by step time series analysis:

See Figure 6.59

Stationary testing:

There are several methods to test stationarity, it include:

1. With White Noise:

White noise is the simplest example of a stationary process. Gausches White Noise is best model because of the following reasons.

- Predictability: If your time series is white noise, then by definition it is random. You can't reasonably model it and make predictions.
- Model diagnosis: The error series of a time series prediction model should ideally be white noise.
- Gauss or Normal distribution is important distribution because of some reasons

To identify white noise, use `ACF` (first identification) and `Ljung-Box-Pierce-Test` (uncorrelation).

```
par(mfrow=c(1,1), mar=c(3,3,1,0)+.5, mgp=c(1.6,.6,0))
acf(sales_ts_ohne_na)
```

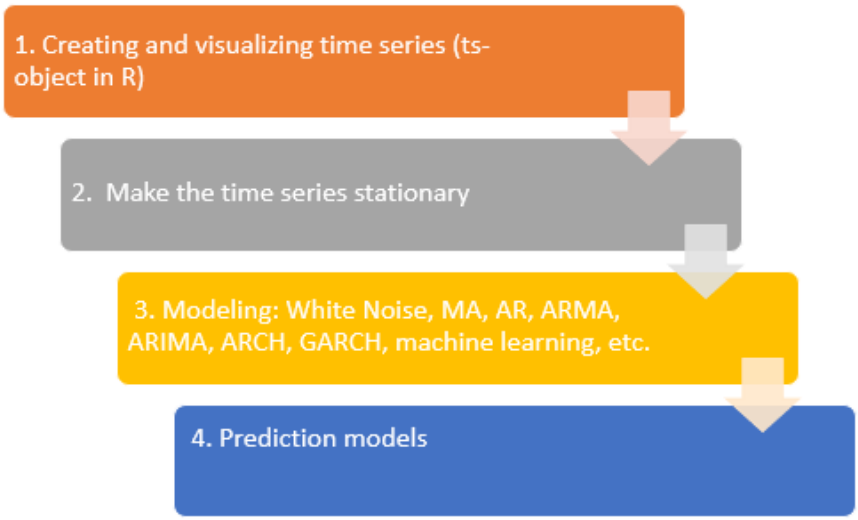
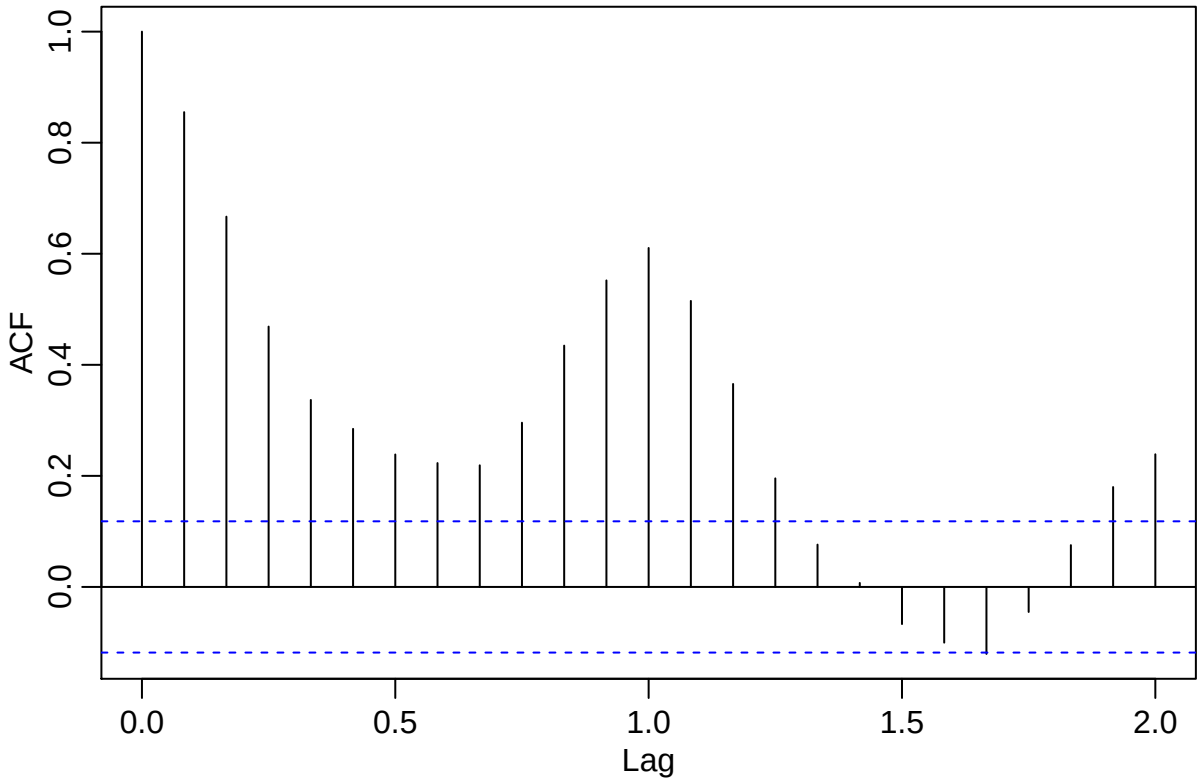

1. Creating and visualizing time series (ts-object in R)
2. Make the time series stationary
3. Modeling: White Noise, MA, AR, ARMA, ARIMA, ARCH, GARCH, machine learning, etc.
4. Prediction models

Figure 6.57:


ACF
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
Lag

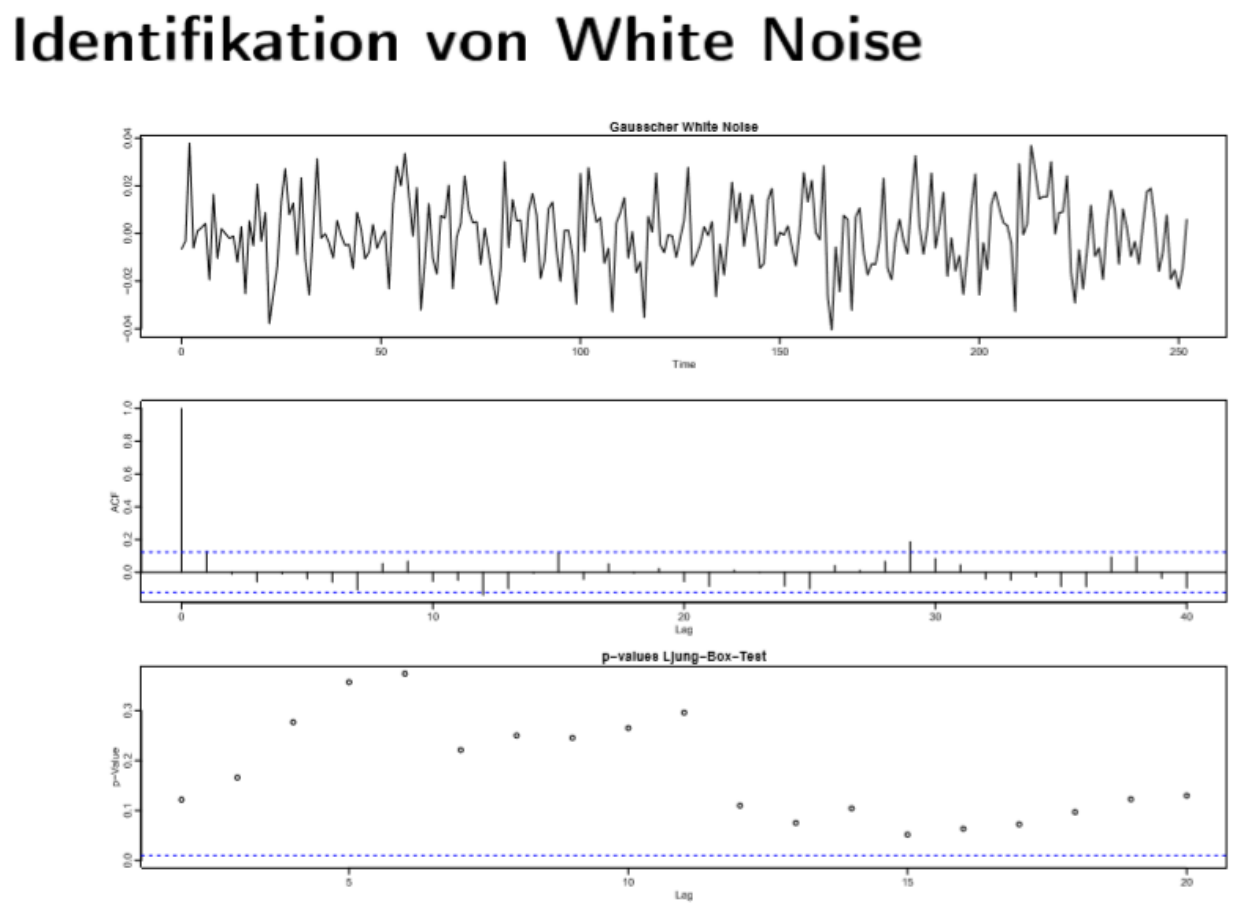


Figure 6.58:

```
Box.test(sales_ts_ohne_na, lag=30, type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  sales_ts_ohne_na
## X-squared = 1025.2, df = 30, p-value < 2.2e-16
```

See Figure 6.60

2. Another Test:

Uncorrelation such as: Augmented Dickey Fuller test, Phillipps-Perron test, Zivot-Andrews test, ADF-GLS test.

```
install.packages("tseries") library(tseries) adf.test(sales_ts_ohne_na)
```

```
# Augmented Dickey-Fuller Test
#
# data:  sales_ts_ohne_na
# Dickey-Fuller = -3.9397, Lag order = 6, p-value = 0.01247
# alternative hypothesis: stationary
```

- Phillipps-Perron test:

```
pp.test(sales_ts_ohne_na)
```

```
# p-value smaller than printed p-value
#   Phillips-Perron Unit Root Test
#
# data:  sales_ts_ohne_na
# Dickey-Fuller Z(alpha) = -44.195, Truncation lag parameter = 5, p-value = 0.01
# alternative hypothesis: stationary
```

Seasonality and trend extraction:

There are some methods to decompose trend or season, it include:

Decompose Trend: -> Trend Residuals= Data - Season

Trend zerlegen: -> Trendresiduen= Data-Season

- Global Trend
  - Linear
  - Nicht linear
- Lokal Trend
  - Einfache gleitende Durchschnitte
  - Differenzfilter
  - Faltung: Hintereinanderausführung von Filtern
- Saison bereinigung: -> Saisonresiduen = Daten - Trend
  - Phasendurchschnittsverfahren
  - Regression mittels Saison-Dummies
  - Regression mit trigonometrischen Polynomen
  - Differenzenbildung
  - Viele etablierte Verfahren als Mischung aus diesen Methoden
- Saison bereinigung: -> Saisonresiduen = Daten - Trend
  - Phasendurchschnittsverfahren
  - Regression mittels Saison-Dummies
  - Regression mit trigonometrischen Polynomen
  - Differenzenbildung
  - Viele etablierte Verfahren als Mischung aus diesen Methoden
- Seasonal adjustment: -> seasonal residuals = data - trend
  - Phase average method
  - Regression using seasonal dummies
  - Regression with trigonometric polynomials
  - Difference formation
  - Many established methods as a mixture of these methods

Direkt: `DECOMPOSE( )` and `STL()`: Time Series Decomposition in R:

To make life easier, some R-packages offer decomposition with a single line of code.

`DECOMPOSE()`:

```
decompose(sales_ts_ohne_na, type = "additive")
```

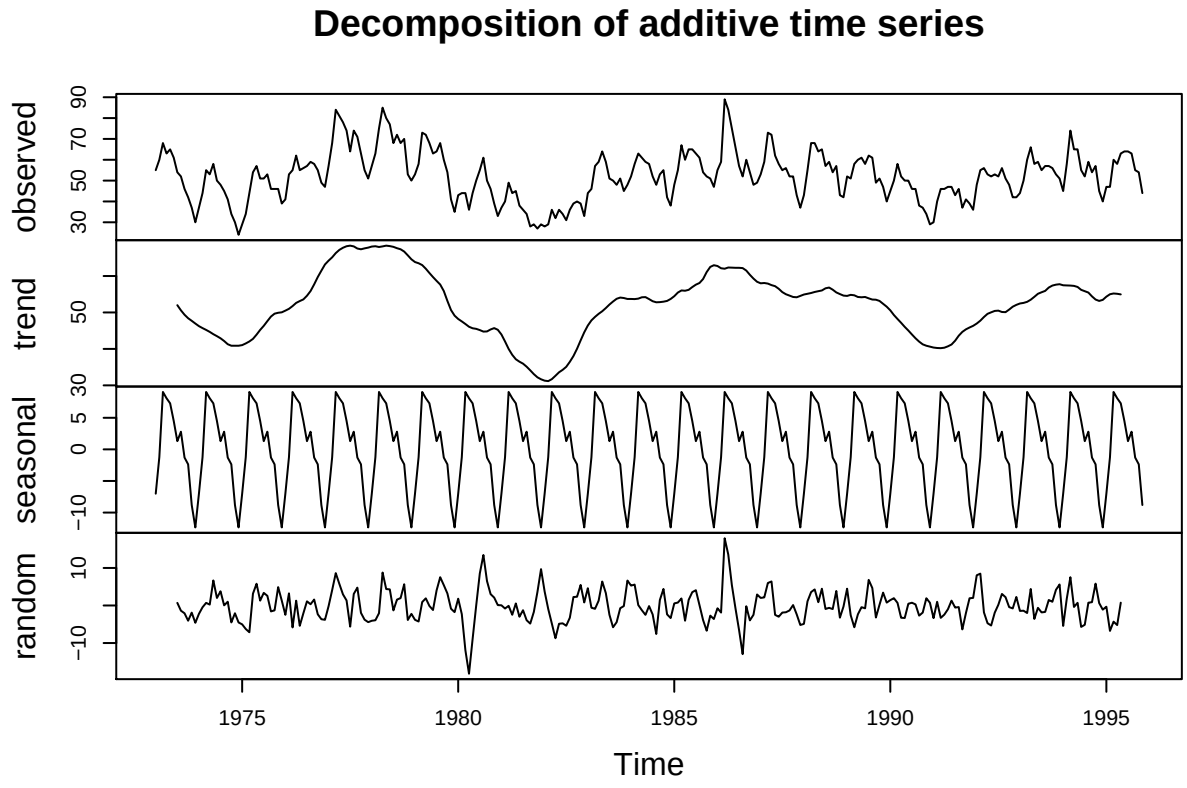
```
## $x
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1973  55  60  68  63  65  61  54  52  46  42  37  30
## 1974  37  44  55  53  58  50  48  45  41  34  30  24
## 1975  29  34  44  54  57  51  51  53  46  46  46  39
## 1976  41  53  55  62  55  56  57  59  58  55  49  47
## 1977  57  68  84  81  78  74  64  74  71  63  55  51
## 1978  57  63  75  85  80  77  68  72  68  70  53  50
## 1979  53  58  73  72  68  63  64  68  60  54  41  35
## 1980  43  44  44  36  44  50  55  61  50  46  39  33
## 1981  37  40  49  44  45  38  36  34  28  29  27  29
## 1982  28  29  36  32  36  34  31  36  39  40  39  33
## 1983  44  46  57  59  64  59  51  50  48  51  45  48
## 1984  52  58  63  61  59  58  52  48  53  55  42  38
## 1985  48  55  67  60  65  65  63  61  54  52  51  47
## 1986  55  59  89  84  75  66  57  52  60  54  48  49
## 1987  53  59  73  72  62  58  55  56  52  52  43  37
## 1988  43  55  68  68  64  65  57  59  54  57  43  42
## 1989  52  51  58  60  61  58  62  61  49  51  47  40
## 1990  45  50  58  52  50  50  46  46  38  37  34  29
## 1991  30  40  46  46  47  47  43  46  37  41  39  36
## 1992  48  55  56  53  52  53  52  56  51  48  42  42
## 1993  44  50  60  66  58  59  55  57  57  56  53  51
## 1994  45  58  74  65  65  55  52  59  54  57  45  40
## 1995  47  47  60  58  63  64  64  63  55  54  44
##
## $seasonal
##              Jan        Feb        Mar        Apr        May        Jun
## 1973  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1974  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1975  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1976  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1977  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1978  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1979  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1980  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1981  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1982  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1983  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1984  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1985  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1986  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1987  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1988  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1989  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1990  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1991  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1992  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
```

```
## 1993  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1994  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
## 1995  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
##             Jul        Aug        Sep        Oct        Nov        Dec
## 1973   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1974   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1975   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1976   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1977   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1978   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1979   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1980   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1981   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1982   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1983   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1984   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1985   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1986   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1987   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1988   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1989   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1990   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1991   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1992   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1993   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1994   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
## 1995   1.304022   2.798341  -1.298250  -2.364538  -8.805826
##
## $trend
##           Jan      Feb      Mar      Apr      May      Jun      Jul
## 1973       NA       NA       NA       NA       NA       NA 52.00000
## 1974 46.25000 45.70833 45.20833 44.66667 44.04167 43.50000 42.91667
## 1975 41.04167 41.50000 42.04167 42.75000 43.91667 45.20833 46.33333
## 1976 50.50000 51.00000 51.75000 52.62500 53.12500 53.58333 54.58333
## 1977 64.20833 65.12500 66.29167 67.16667 67.75000 68.16667 68.33333
## 1978 68.08333 68.16667 67.95833 68.12500 68.33333 68.20833 68.00000
## 1979 63.83333 63.50000 63.00000 62.00000 60.83333 59.70833 58.66667
## 1980 48.20833 47.54167 46.83333 46.08333 45.66667 45.50000 45.16667
## 1981 43.95833 42.04167 40.00000 38.37500 37.16667 36.50000 35.95833
## 1982 31.29167 31.16667 31.70833 32.62500 33.58333 34.25000 35.08333
## 1983 46.41667 47.83333 48.79167 49.62500 50.33333 51.20833 52.16667
## 1984 53.70833 53.66667 53.79167 54.16667 54.20833 53.66667 53.08333
## 1985 54.45833 55.45833 56.04167 55.95833 56.20833 56.95833 57.62500
## 1986 62.75000 62.12500 62.00000 62.33333 62.29167 62.25000 62.25000
## 1987 58.00000 58.08333 57.91667 57.50000 57.20833 56.50000 55.58333
## 1988 54.91667 55.12500 55.33333 55.62500 55.83333 56.04167 56.62500
## 1989 54.54167 54.83333 54.70833 54.25000 54.16667 54.25000 53.87500
## 1990 50.58333 49.29167 48.20833 47.16667 46.04167 45.04167 43.95833
## 1991 40.37500 40.25000 40.20833 40.33333 40.70833 41.20833 42.25000
## 1992 46.95833 47.75000 48.75000 49.62500 50.04167 50.41667 50.50000
## 1993 52.45833 52.62500 52.91667 53.50000 54.29167 55.12500 55.54167
## 1994 57.45833 57.41667 57.37500 57.29167 57.00000 56.20833 55.83333
## 1995 54.33333 55.00000 55.20833 55.12500 54.95833       NA       NA
##           Aug      Sep      Oct      Nov      Dec
```

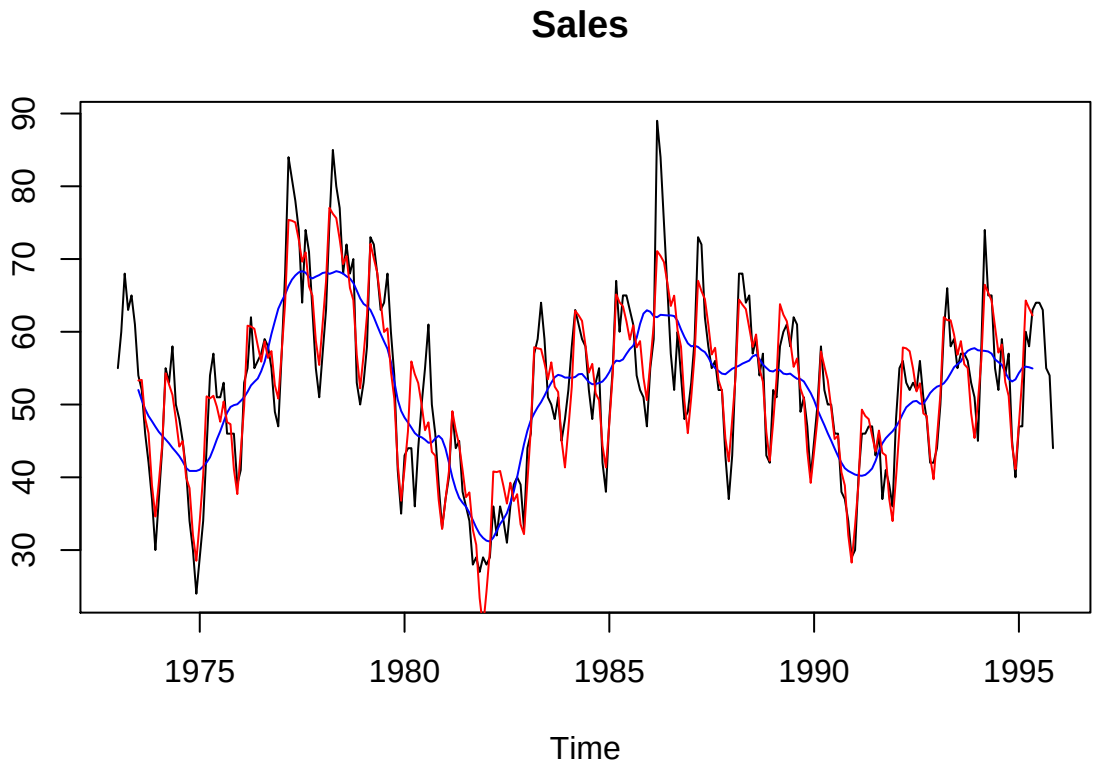
```
## 1973 50.58333 49.37500 48.41667 47.70833 46.95833
## 1974 42.16667 41.29167 40.87500 40.87500 40.87500
## 1975 47.62500 48.87500 49.66667 49.91667 50.04167
## 1976 55.87500 57.70833 59.70833 61.45833 63.16667
## 1977 68.12500 67.54167 67.33333 67.58333 67.79167
## 1978 67.62500 67.33333 66.70833 65.66667 64.58333
## 1979 57.66667 55.87500 53.16667 50.66667 49.12500
## 1980 44.75000 44.79167 45.33333 45.70833 45.25000
## 1981 35.12500 34.12500 33.08333 32.20833 31.66667
## 1982 36.45833 38.04167 40.04167 42.33333 44.54167
## 1983 53.00000 53.75000 54.08333 53.95833 53.70833
## 1984 52.79167 52.83333 52.95833 53.16667 53.70833
## 1985 58.08333 59.16667 61.08333 62.50000 62.95833
## 1986 62.16667 61.50000 60.33333 59.29167 58.41667
## 1987 55.00000 54.62500 54.25000 54.16667 54.54167
## 1988 56.83333 56.25000 55.50000 55.04167 54.62500
## 1989 53.54167 53.50000 53.16667 52.37500 51.58333
## 1990 42.91667 42.00000 41.25000 40.87500 40.62500
## 1991 43.62500 44.66667 45.37500 45.87500 46.33333
## 1992 50.12500 50.08333 50.79167 51.58333 52.08333
## 1993 55.91667 56.83333 57.37500 57.62500 57.75000
## 1994 55.45833 54.41667 53.54167 53.16667 53.45833
## 1995       NA       NA       NA       NA
##
## $random
##                Jan           Feb           Mar           Apr           May
## 1973            NA            NA            NA            NA            NA
## 1974  -2.235840548  -0.472582973   0.701659452   0.237644300   6.671356421
## 1975  -5.027507215  -6.264249639  -7.131673882   3.154310967   5.796356421
## 1976  -2.485840548   3.235750361  -5.840007215   1.279310967  -5.411976912
## 1977  -0.194173882   4.110750361   8.618326118   5.737644300   2.963023088
## 1978  -4.069173882  -3.930916306  -2.048340548   8.779310967   4.379689755
## 1979  -3.819173882  -4.264249639   0.909992785   1.904310967  -0.120310245
## 1980   1.805826118  -2.305916306 -11.923340548 -18.179022367  -8.953643579
## 1981   0.055826118  -0.805916306  -0.090007215  -2.470689033   0.546356421
## 1982   3.722492785  -0.930916306  -4.798340548  -8.720689033  -4.870310245
## 1983   4.597492785  -0.597582973  -0.881673882   1.279310967   6.379689755
## 1984   5.305826118   5.569083694   0.118326118  -1.262355700  -2.495310245
## 1985   0.555826118   0.777417027   1.868326118  -4.054022367   1.504689755
## 1986  -0.735840548  -1.889249639  17.909992785  13.570977633   5.421356421
## 1987   2.014159452   2.152417027   5.993326118   6.404310967  -2.495310245
## 1988  -4.902507215   1.110750361   3.576659452   4.279310967   0.879689755
## 1989   4.472492785  -2.597582973  -5.798340548  -2.345689033  -0.453643579
## 1990   1.430826118   1.944083694   0.701659452  -3.262355700  -3.328643579
## 1991  -3.360840548   0.985750361  -3.298340548  -2.429022367  -0.995310245
## 1992   8.055826118   8.485750361  -1.840007215  -4.720689033  -5.328643579
## 1993  -1.444173882  -1.389249639  -2.006673882   4.404310967  -3.578643579
## 1994  -5.444173882   1.819083694   7.534992785  -0.387355700   0.713023088
## 1995  -0.319173882  -6.764249639  -4.298340548  -5.220689033   0.754689755
##                Jun           Jul           Aug           Sep           Oct
## 1973            NA   0.695977633  -1.381673882  -2.076749639  -4.052128427
## 1974   2.003336941   3.779310967   0.034992785   1.006583694  -4.510461760
## 1975   1.295003608   3.362644300   2.576659452  -1.576749639  -1.302128427
## 1976  -2.079996392   1.112644300   0.326659452   1.589917027  -2.343795094
```

```
## 1977   1.336670274  -5.637355700   3.076659452   4.756583694  -1.968795094
## 1978   4.295003608  -1.304022367   1.576659452   1.964917027   5.656204906
## 1979  -1.204996392   4.029310967   7.534992785   5.423250361   3.197871573
## 1980   0.003336941   8.529310967  13.451659452   6.506583694   3.031204906
## 1981  -2.996663059  -1.262355700  -3.923340548  -4.826749639  -1.718795094
## 1982  -4.746663059  -5.387355700  -3.256673882   2.256583694   2.322871573
## 1983   3.295003608  -2.470689033  -5.798340548  -4.451749639  -0.718795094
## 1984  -0.163329726  -2.387355700  -7.590007215   1.464917027   4.406204906
## 1985   3.545003608   4.070977633   0.118326118  -3.868416306  -6.718795094
## 1986  -0.746663059  -6.554022367 -12.965007215  -0.201749639  -3.968795094
## 1987  -2.996663059  -1.887355700  -1.798340548  -1.326749639   0.114538240
## 1988   4.461670274  -0.929022367  -0.631673882  -0.951749639   3.864538240
## 1989  -0.746663059   6.820977633   4.659992785  -3.201749639   0.197871573
## 1990   0.461670274   0.737644300   0.284992785  -2.701749639  -1.885461760
## 1991   1.295003608  -0.554022367  -0.423340548  -6.368416306  -2.010461760
## 1992  -1.913329726   0.195977633   3.076659452   2.214917027  -0.427128427
## 1993  -0.621663059  -1.845689033  -1.715007215   1.464917027   0.989538240
## 1994  -5.704996392  -5.137355700   0.743326118   0.881583694   5.822871573
## 1995            NA            NA            NA            NA            NA
##               Nov           Dec
## 1973  -1.902507215  -4.605158730
## 1974  -2.069173882  -4.521825397
## 1975   4.889159452   1.311507937
## 1976  -3.652507215  -3.813492063
## 1977  -3.777507215  -4.438492063
## 1978  -3.860840548  -2.230158730
## 1979  -0.860840548  -1.771825397
## 1980   2.097492785   0.103174603
## 1981   3.597492785   9.686507937
## 1982   5.472492785   0.811507937
## 1983  -0.152507215   6.644841270
## 1984  -2.360840548  -3.355158730
## 1985  -2.694173882  -3.605158730
## 1986  -2.485840548   2.936507937
## 1987  -2.360840548  -5.188492063
## 1988  -3.235840548  -0.271825397
## 1989   3.430826118   0.769841270
## 1990   1.930826118   0.728174603
## 1991   1.930826118   2.019841270
## 1992  -0.777507215   2.269841270
## 1993   4.180826118   5.603174603
## 1994   0.639159452  -1.105158730
## 1995            NA
##
## $figure
##  [1]  -7.014159  -1.235750   9.090007   8.095689   7.286977   4.496663
##  [7]   1.304022   2.798341  -1.298250  -2.364538  -8.805826 -12.353175
##
## $type
## [1] "additive"
##
## attr(,"class")
## [1] "decomposed.ts"
```

```
dec=decompose(sales_ts_ohne_na)
plot(dec)
```




```
plot(sales_ts_ohne_na, main="Sales ", ylab="")
lines((dec$trend), col=4)
lines((dec$seasonal+dec$trend), col=2)
```



STL():

```
par(mfrow=c(1,1))
dec.pass=stl(sales_ts_ohne_na,"periodic")
dec.pass
```

```
##  Call:
##  stl(x = sales_ts_ohne_na, s.window = "periodic")
##
## Components
##             seasonal    trend    remainder
## Jan 1973  -7.052492 61.82257   0.22992302
## Feb 1973  -1.230382 60.22258   1.00779851
## Mar 1973   9.069991 58.62260   0.30741135
## Apr 1973   7.967740 57.06243  -2.03016662
## May 1973   7.343751 55.50225   2.15399468
## Jun 1973   4.752665 54.00072   2.24661366
## Jul 1973   1.552883 52.49919  -0.05207051
## Aug 1973   2.915340 51.05776  -1.97309725
## Sep 1973  -1.374374 49.61633  -2.24195230
## Oct 1973  -2.458432 48.47375  -4.01531472
## Nov 1973  -9.064229 47.33117  -1.26693669
## Dec 1973 -12.422461 46.71048  -4.28801925
## Jan 1974  -7.052492 46.08979  -2.03730284
## Feb 1974  -1.230382 45.68398  -0.45359851
## Mar 1974   9.069991 45.27817   0.65184318
## Apr 1974   7.967740 44.83067   0.20159186
```

```
## May 1974   7.343751 44.38317   6.27307982
## Jun 1974   4.752665 43.73419   1.51314832
## Jul 1974   1.552883 43.08520   3.36191367
## Aug 1974   2.915340 42.27816  -0.19349857
## Sep 1974  -1.374374 41.47111   0.90326087
## Oct 1974  -2.458432 40.95995  -4.50151810
## Nov 1974  -9.064229 40.44879  -1.38455662
## Dec 1974 -12.422461 40.52756  -4.10509649
## Jan 1975  -7.052492 40.60633  -4.55383740
## Feb 1975  -1.230382 41.25403  -6.02364619
## Mar 1975   9.069991 41.90173  -6.97171763
## Apr 1975   7.967740 42.99659   3.03567358
## May 1975   7.343751 44.09145   5.56480408
## Jun 1975   4.752665 45.38910   0.85823284
## Jul 1975   1.552883 46.68676   2.76035846
## Aug 1975   2.915340 47.76362   2.32103732
## Sep 1975  -1.374374 48.84049  -1.46611212
## Oct 1975  -2.458432 49.39289  -0.93446019
## Nov 1975  -9.064229 49.94530   5.11893219
## Dec 1975 -12.422461 50.26758   1.15487730
## Jan 1976  -7.052492 50.58987  -2.53737863
## Feb 1976  -1.230382 51.09301   3.13737254
## Mar 1976   9.069991 51.59615  -5.66613893
## Apr 1976   7.967740 52.24755   1.78470749
## May 1976   7.343751 52.89896  -5.24270682
## Jun 1976   4.752665 53.78671  -2.53937448
## Jul 1976   1.552883 54.67446   0.77265471
## Aug 1976   2.915340 56.14859  -0.06392721
## Sep 1976  -1.374374 57.62271   1.75166257
## Oct 1976  -2.458432 59.43175  -1.97331844
## Nov 1976  -9.064229 61.24079  -3.17655900
## Dec 1976 -12.422461 62.77446  -3.35199618
## Jan 1977  -7.052492 64.30813  -0.25563440
## Feb 1977  -1.230382 65.44889   3.78148986
## Mar 1977   9.069991 66.58966   8.34035148
## Apr 1977   7.967740 67.32774   5.70452219
## May 1977   7.343751 68.06582   2.59043218
## Jun 1977   4.752665 68.18920   1.05813960
## Jul 1977   1.552883 68.31257  -5.86545612
## Aug 1977   2.915340 67.95156   3.13310168
## Sep 1977  -1.374374 67.59054   4.78383117
## Oct 1977  -2.458432 67.45458  -1.99615263
## Nov 1977  -9.064229 67.31863  -3.25439598
## Dec 1977 -12.422461 67.50216  -4.07970363
## Jan 1978  -7.052492 67.68570  -3.63321231
## Feb 1978  -1.230382 67.90944  -3.67905629
## Mar 1978   9.069991 68.13317  -2.20316292
## Apr 1978   7.967740 68.31711   8.71514611
## May 1978   7.343751 68.50105   4.15519441
## Jun 1978   4.752665 68.43810   3.80923053
## Jul 1978   1.552883 68.37515  -1.92803649
## Aug 1978   2.915340 67.82881   1.25584535
## Sep 1978  -1.374374 67.28248   2.09189888
## Oct 1978  -2.458432 66.41100   6.04742871
```

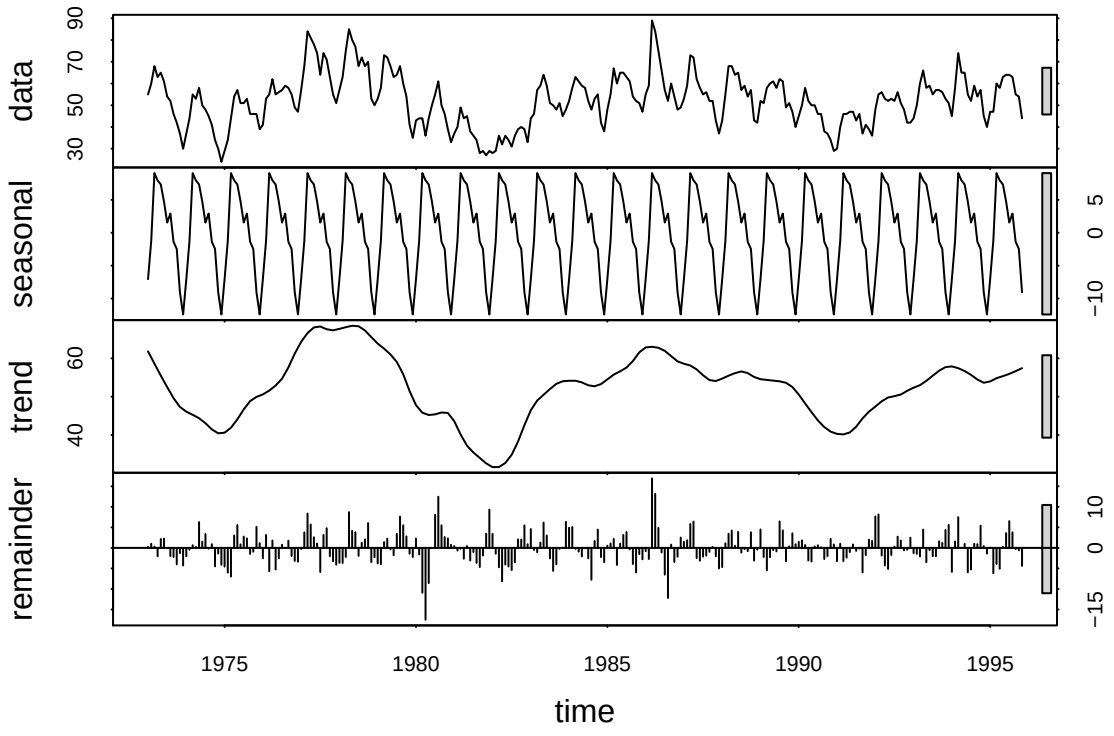
```
## Nov 1978  -9.064229 65.53953  -3.47530102
## Dec 1978 -12.422461 64.69248  -2.27002103
## Jan 1979  -7.052492 63.84543  -3.79294208
## Feb 1979  -1.230382 63.18734  -3.95696255
## Mar 1979   9.069991 62.52925   1.40075433
## Apr 1979   7.967740 61.77447   2.25778892
## May 1979   7.343751 61.01969  -0.36343721
## Jun 1979   4.752665 60.03600  -1.78866950
## Jul 1979   1.552883 59.05232   3.39479508
## Aug 1979   2.915340 57.45553   7.62912815
## Sep 1979  -1.374374 55.85874   5.51563292
## Oct 1979  -2.458432 53.66388   2.79454913
## Nov 1979  -9.064229 51.46902  -1.40479421
## Dec 1979 -12.422461 49.63182  -2.20935852
## Jan 1980  -7.052492 47.79462   2.25787613
## Feb 1980  -1.230382 46.81209  -1.58171047
## Mar 1980   9.069991 45.82957 -10.89955971
## Apr 1980   7.967740 45.52741 -17.49515297
## May 1980   7.343751 45.22526  -8.56900695
## Jun 1980   4.752665 45.30657  -0.05923179
## Jul 1980   1.552883 45.38788   8.05924023
## Aug 1980   2.915340 45.63098  12.45367875
## Sep 1980  -1.374374 45.87409   5.50028896
## Oct 1980  -2.458432 45.81279   2.64563820
## Nov 1980  -9.064229 45.75150   2.31272788
## Dec 1980 -12.422461 44.67895   0.74351401
## Jan 1981  -7.052492 43.60639   0.44609911
## Feb 1981  -1.230382 41.86292  -0.63254021
## Mar 1981   9.069991 40.11945  -0.18944217
## Apr 1981   7.967740 38.66696  -2.63470542
## May 1981   7.343751 37.21448   0.44177061
## Jun 1981   4.752665 36.32708  -3.07974602
## Jul 1981   1.552883 35.43968  -0.99256579
## Aug 1981   2.915340 34.73098  -3.64631921
## Sep 1981  -1.374374 34.02228  -4.64790094
## Oct 1981  -2.458432 33.31112  -1.85268400
## Nov 1981  -9.064229 32.59996   3.46427338
## Dec 1981 -12.422461 32.11584   9.30662189
## Jan 1982  -7.052492 31.63172   3.42076936
## Feb 1982  -1.230382 31.63684  -1.40645631
## Mar 1982   9.069991 31.64195  -4.71194463
## Apr 1982   7.967740 32.17310  -8.14083575
## May 1982   7.343751 32.70424  -4.04798760
## Jun 1982   4.752665 33.77989  -4.53255552
## Jul 1982   1.552883 34.85554  -5.40842659
## Aug 1982   2.915340 36.57721  -3.49255085
## Sep 1982  -1.374374 38.29888   2.07549658
## Oct 1982  -2.458432 40.44169   2.01673668
## Nov 1982  -9.064229 42.58451   5.47971723
## Dec 1982 -12.422461 44.52271   0.89975396
## Jan 1983  -7.052492 46.46090   4.59158966
## Feb 1983  -1.230382 47.72738  -0.49700185
## Mar 1983   9.069991 48.99387  -1.06385600
## Apr 1983   7.967740 49.74380   1.28846446
```

```
## May 1983   7.343751 50.49373   6.16252419
## Jun 1983   4.752665 51.24299   3.00434902
## Jul 1983   1.552883 51.99225  -2.54512929
## Aug 1983   2.915340 52.67312  -5.58846331
## Sep 1983  -1.374374 53.35400  -3.97962563
## Oct 1983  -2.458432 53.68724  -0.22880872
## Nov 1983  -9.064229 54.02048   0.04374863
## Dec 1983 -12.422461 54.08807   6.33439123
## Jan 1984  -7.052492 54.15566   4.89683279
## Feb 1984  -1.230382 54.15654   5.07383944
## Mar 1984   9.069991 54.15743  -0.22741656
## Apr 1984   7.967740 53.94335  -0.91108638
## May 1984   7.343751 53.72927  -2.07301693
## Jun 1984   4.752665 53.35159  -0.10425199
## Jul 1984   1.552883 52.97391  -2.52679020
## Aug 1984   2.915340 52.84330  -7.75864272
## Sep 1984  -1.374374 52.71270   1.66167645
## Oct 1984  -2.458432 52.99779   4.46064056
## Nov 1984  -9.064229 53.28288  -2.21865489
## Dec 1984 -12.422461 53.89935  -3.47688432
## Jan 1985  -7.052492 54.51581   0.53668521
## Feb 1985  -1.230382 55.13482   1.09556329
## Mar 1985   9.069991 55.75383   2.17617872
## Apr 1985   7.967740 56.18577  -4.15350757
## May 1985   7.343751 56.61770   1.03854542
## Jun 1985   4.752665 57.10283   3.14450950
## Jul 1985   1.552883 57.58795   3.85917043
## Aug 1985   2.915340 58.43915  -0.35449122
## Sep 1985  -1.374374 59.29036  -3.91598117
## Oct 1985  -2.458432 60.42009  -5.96166073
## Nov 1985  -9.064229 61.54983  -1.48559984
## Dec 1985 -12.422461 62.19821  -2.77574578
## Jan 1986  -7.052492 62.84658  -0.79409276
## Feb 1986  -1.230382 62.92917  -2.69879287
## Mar 1986   9.069991 63.01176  16.91824438
## Apr 1986   7.967740 62.88496  13.14729507
## May 1986   7.343751 62.75816   4.89808503
## Jun 1986   4.752665 62.35540  -1.10806094
## Jul 1986   1.552883 61.95263  -6.50551005
## Aug 1986   2.915340 61.26809 -12.18343198
## Sep 1986  -1.374374 60.58356   0.79081778
## Oct 1986  -2.458432 59.90096  -3.44252369
## Nov 1986  -9.064229 59.21835  -2.15412471
## Dec 1986 -12.422461 58.90597   2.51649340
## Jan 1987  -7.052492 58.59358   1.45891048
## Feb 1987  -1.230382 58.36275   1.86763567
## Mar 1987   9.069991 58.13191   5.79809822
## Apr 1987   7.967740 57.62558   6.40668384
## May 1987   7.343751 57.11924  -2.46299126
## Jun 1987   4.752665 56.35507  -3.10773342
## Jul 1987   1.552883 55.59090  -2.14377872
## Aug 1987   2.915340 54.96140  -1.87673866
## Sep 1987  -1.374374 54.33190  -0.95752691
## Oct 1987  -2.458432 54.20957   0.24886271
```

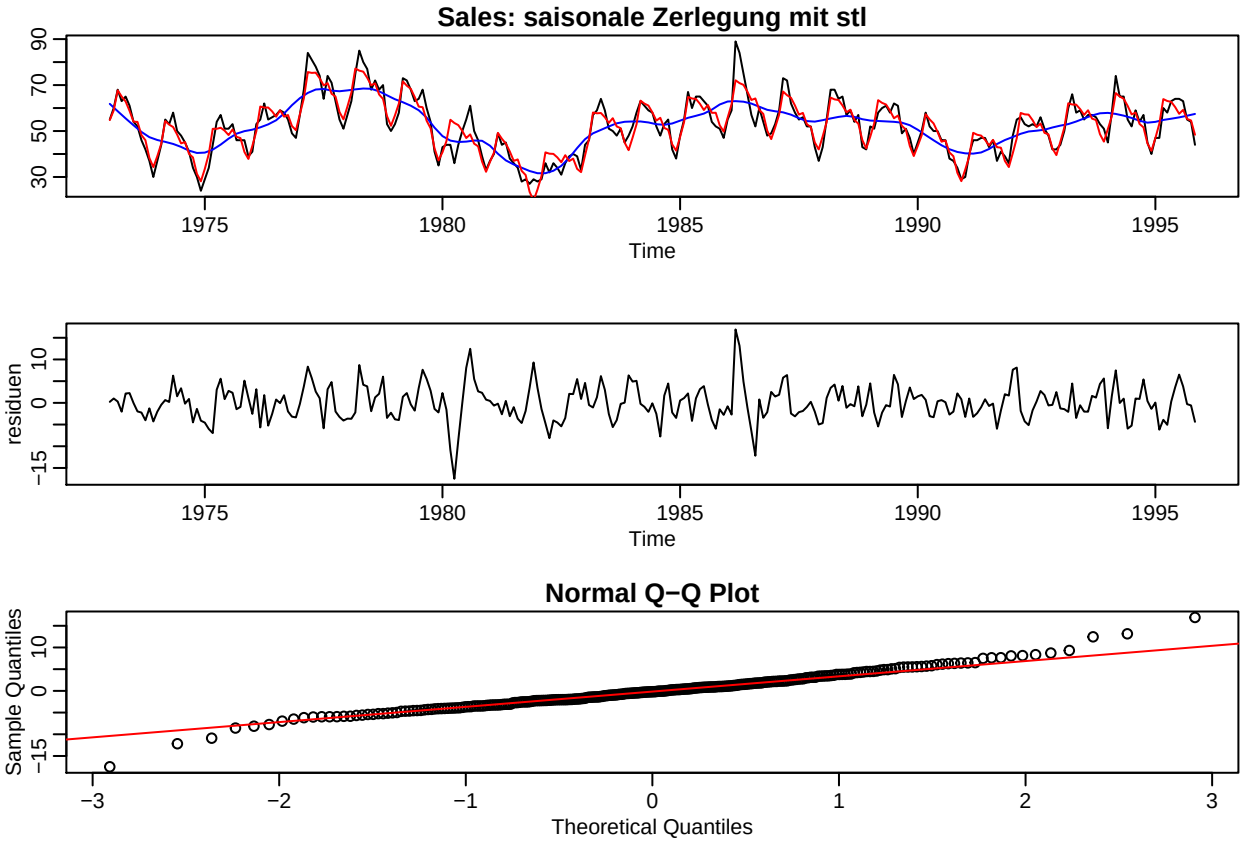
```
## Nov 1987  -9.064229 54.08724  -2.02300724
## Dec 1987 -12.422461 54.40582  -4.98335841
## Jan 1988  -7.052492 54.72440  -4.67191061
## Feb 1988  -1.230382 55.09893   1.13145559
## Mar 1988   9.069991 55.47345   3.45655914
## Apr 1988   7.967740 55.80992   4.22234222
## May 1988   7.343751 56.14638   0.50986457
## Jun 1988   4.752665 56.36676   3.88057176
## Jul 1988   1.552883 56.58714  -1.14002420
## Aug 1988   2.915340 56.37484  -0.29018081
## Sep 1988  -1.374374 56.16254  -0.78816573
## Oct 1988  -2.458432 55.65706   3.80136780
## Nov 1988  -9.064229 55.15159  -3.08735824
## Dec 1988 -12.422461 54.85661  -0.43414877
## Jan 1989  -7.052492 54.56163   4.49085967
## Feb 1989  -1.230382 54.46603  -2.23564647
## Mar 1989   9.069991 54.37042  -5.44041525
## Apr 1989   7.967740 54.27968  -2.24741646
## May 1989   7.343751 54.18893  -0.53267839
## Jun 1989   4.752665 54.09677  -0.84943702
## Jul 1989   1.552883 54.00462   6.44250120
## Aug 1989   2.915340 53.81896   4.26570373
## Sep 1989  -1.374374 53.63330  -3.25892204
## Oct 1989  -2.458432 53.07898   0.37945619
## Nov 1989  -9.064229 52.52465   3.53957488
## Dec 1989 -12.422461 51.54381   0.87864696
## Jan 1990  -7.052492 50.56297   1.48951801
## Feb 1990  -1.230382 49.39765   1.83273208
## Mar 1990   9.069991 48.23233   0.69768350
## Apr 1990   7.967740 47.09863  -3.06637382
## May 1990   7.343751 45.96494  -3.30869187
## Jun 1990   4.752665 44.88695   0.36038472
## Jul 1990   1.552883 43.80896   0.63815816
## Aug 1990   2.915340 42.94279   0.14187147
## Sep 1990  -1.374374 42.07662  -2.70224354
## Oct 1990  -2.458432 41.48288  -2.02444702
## Nov 1990  -9.064229 40.88914   2.17508995
## Dec 1990 -12.422461 40.59162   0.83084111
## Jan 1991  -7.052492 40.29410  -3.24160876
## Feb 1991  -1.230382 40.22752   1.00285985
## Mar 1991   9.069991 40.16094  -3.23093419
## Apr 1991   7.967740 40.41417  -2.38191515
## May 1991   7.343751 40.66741  -1.01115683
## Jun 1991   4.752665 41.41709   0.83024231
## Jul 1991   1.552883 42.16678  -0.71966169
## Aug 1991   2.915340 43.25418  -0.16951589
## Sep 1991  -1.374374 44.34157  -5.96719840
## Oct 1991  -2.458432 45.20791  -1.74947505
## Nov 1991  -9.064229 46.07424   1.98998874
## Dec 1991 -12.422461 46.74221   1.68025491
## Jan 1992  -7.052492 47.41017   7.64232004
## Feb 1992  -1.230382 48.10287   8.12751590
## Mar 1992   9.069991 48.79556  -1.86555089
## Apr 1992   7.967740 49.29445  -4.26219411
```

```
## May 1992   7.343751 49.79335  -5.13709804
## Jun 1992   4.752665 49.96677  -1.71943574
## Jul 1992   1.552883 50.14019   0.30692343
## Aug 1992   2.915340 50.35278   2.73188191
## Sep 1992  -1.374374 50.56536   1.80901208
## Oct 1992  -2.458432 51.04985  -0.59142160
## Nov 1992  -9.064229 51.53434  -0.47011483
## Dec 1992 -12.422461 51.94083   2.48162756
## Jan 1993  -7.052492 52.34732  -1.29483108
## Feb 1993  -1.230382 52.67858  -1.44820132
## Mar 1993   9.069991 53.00984  -2.07983421
## Apr 1993   7.967740 53.57410   4.45816441
## May 1993   7.343751 54.13835  -3.48209769
## Jun 1993   4.752665 54.80254  -0.55520124
## Jul 1993   1.552883 55.46673  -2.01960793
## Aug 1993   2.915340 56.11835  -2.03368553
## Sep 1993  -1.374374 56.76997   1.60440857
## Oct 1993  -2.458432 57.24291   1.21551945
## Nov 1993  -9.064229 57.71586   4.34837078
## Dec 1993 -12.422461 57.80625   5.61621212
## Jan 1994  -7.052492 57.89664  -5.84414758
## Feb 1994  -1.230382 57.66865   1.56173466
## Mar 1994   9.069991 57.44065   7.48935426
## Apr 1994   7.967740 57.06520  -0.03294373
## May 1994   7.343751 56.68975   0.96649756
## Jun 1994   4.752665 56.19324  -5.94590874
## Jul 1994   1.552883 55.69674  -5.24961819
## Aug 1994   2.915340 55.08499   0.99966537
## Sep 1994  -1.374374 54.47325   0.90112062
## Oct 1994  -2.458432 54.05382   5.40460953
## Nov 1994  -9.064229 53.63439   0.42983887
## Dec 1994 -12.422461 53.80952  -1.38705781
## Jan 1995  -7.052492 53.98465   0.06784448
## Feb 1995  -1.230382 54.40554  -6.17515534
## Mar 1995   9.069991 54.82643  -3.89641782
## Apr 1995   7.967740 55.09259  -5.06033019
## May 1995   7.343751 55.35875   0.29749670
## Jun 1995   4.752665 55.64907   3.59826236
## Jul 1995   1.552883 55.93939   6.50772488
## Aug 1995   2.915340 56.30171   3.78295467
## Sep 1995  -1.374374 56.66402  -0.28964385
## Oct 1995  -2.458432 57.05055  -0.59211603
## Nov 1995  -9.064229 57.43708  -4.37284777
```

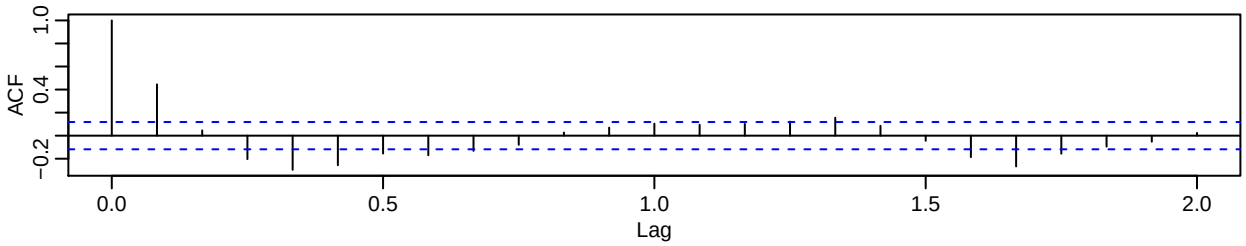
```
plot(dec.pass)
```




```
par(mfrow=c(3,1), mar=c(3,2,1,0)+.5, mgp=c(1.6,.6,0))
plot(sales_ts_ohne_na, main="Sales: saisonale Zerlegung mit stl", ylab="")
lines((dec.pass$time.series[,2]), col=4)
lines((dec.pass$time.series[,2]+dec.pass$time.series[,1]), col=2)



#Residual
residuen=dec.pass$time.series[,3]
plot(residuen)
qqnorm(residuen)
qqline(residuen, col=2)
```




```
acf(residuen)

#Rquadrat(exp(dec.pass$time.series[,2]+dec.pass$time.series[,1]), passb) #Bestimmheitmaß
```



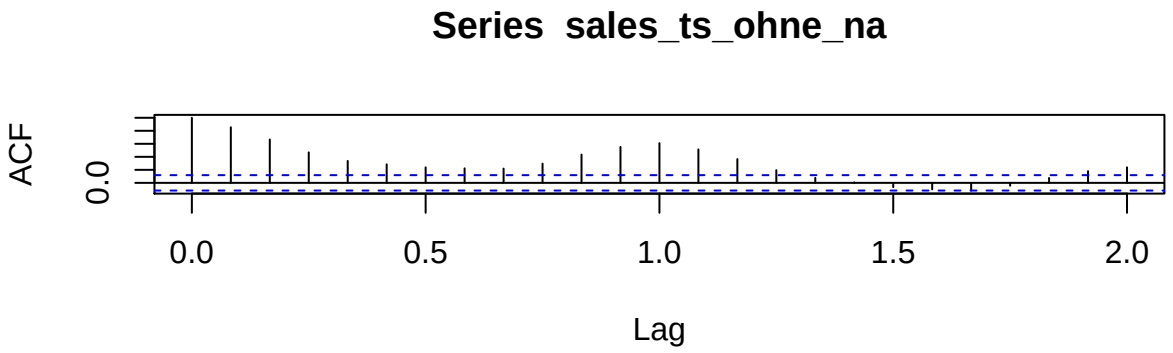
6.5.4 Models

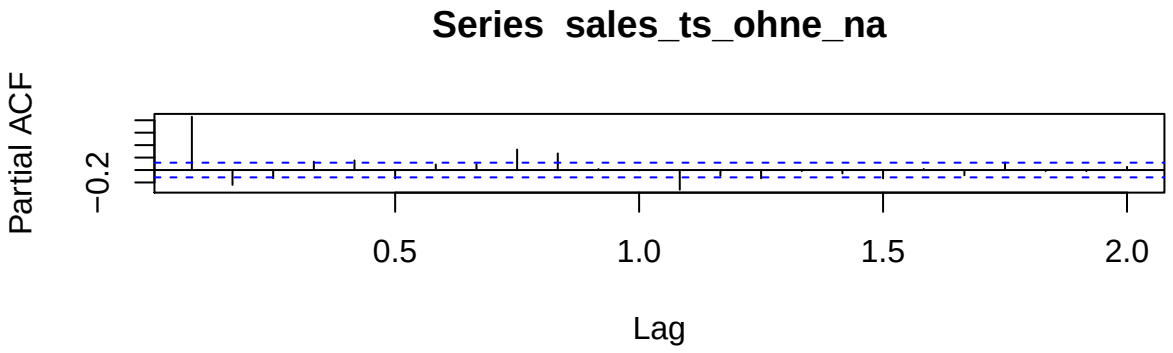
AR Model:

1. ACF and PACF as first identification:

- ACF: exponentially decaying (exponential abklingelnd)
- PACF: order of the models: the PACF has 95% confidence interval

```
par(mfrow=c(2,1))
acf(sales_ts_ohne_na)
pacf(sales_ts_ohne_na)
```





Candicate of orders: 1, 2, 3,4,5….etc

2. Estimate AR parameters:

- Conditional bzw. Unconditional Least Sum of Squares (CLS bzw. ULS)
- Conditional bzw. Unconditional Maximum Likelihood (CML bzw. UML)
- Burg
- Yule Walker (YW)

```
#find the order
mm1_mle=ar.mle(sales_ts_ohne_na)
mm1_mle$order # Ordnung suchen
```

```
## [1] 10
```

```
mm2_yw=ar.yw(sales_ts_ohne_na)
mm2_yw$order # Ordnung suchen
```

```
## [1] 15
```

```
mm3_ols=ar.ols(sales_ts_ohne_na)
mm3_ols$order # Ordnung suchen
```

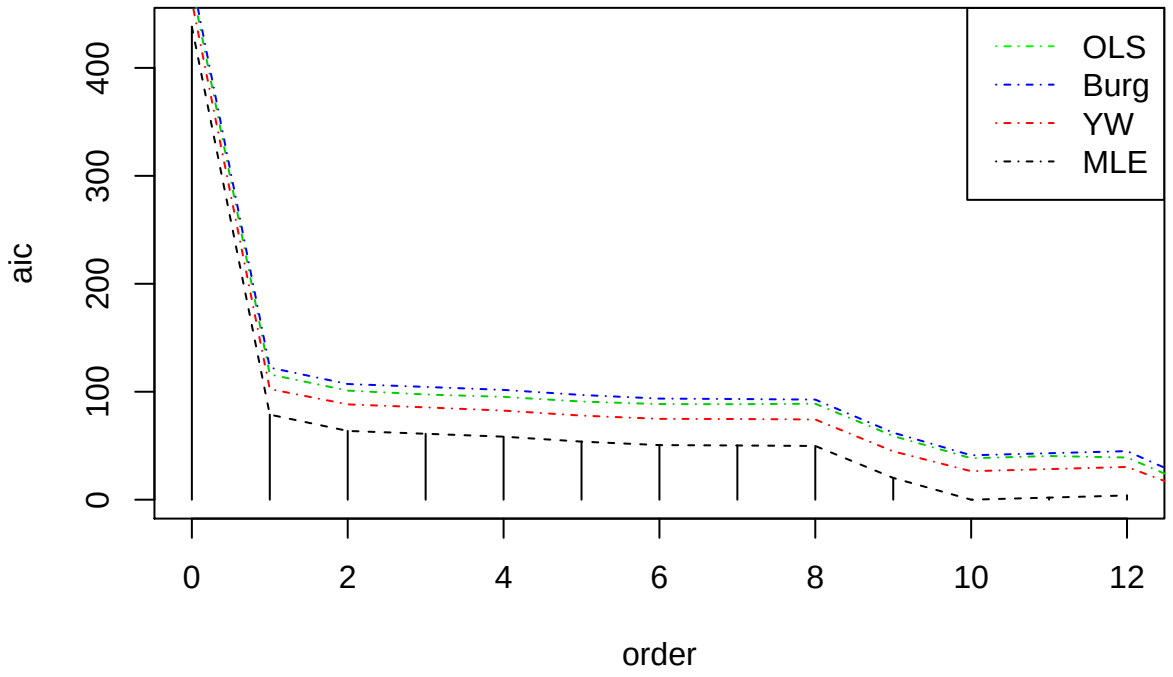
```
## [1] 21
```

```
mm4_burg=ar.burg(sales_ts_ohne_na)
mm4_burg$order # Ordnung suchen
```

```
## [1] 21
```

Graphical comparison methods:

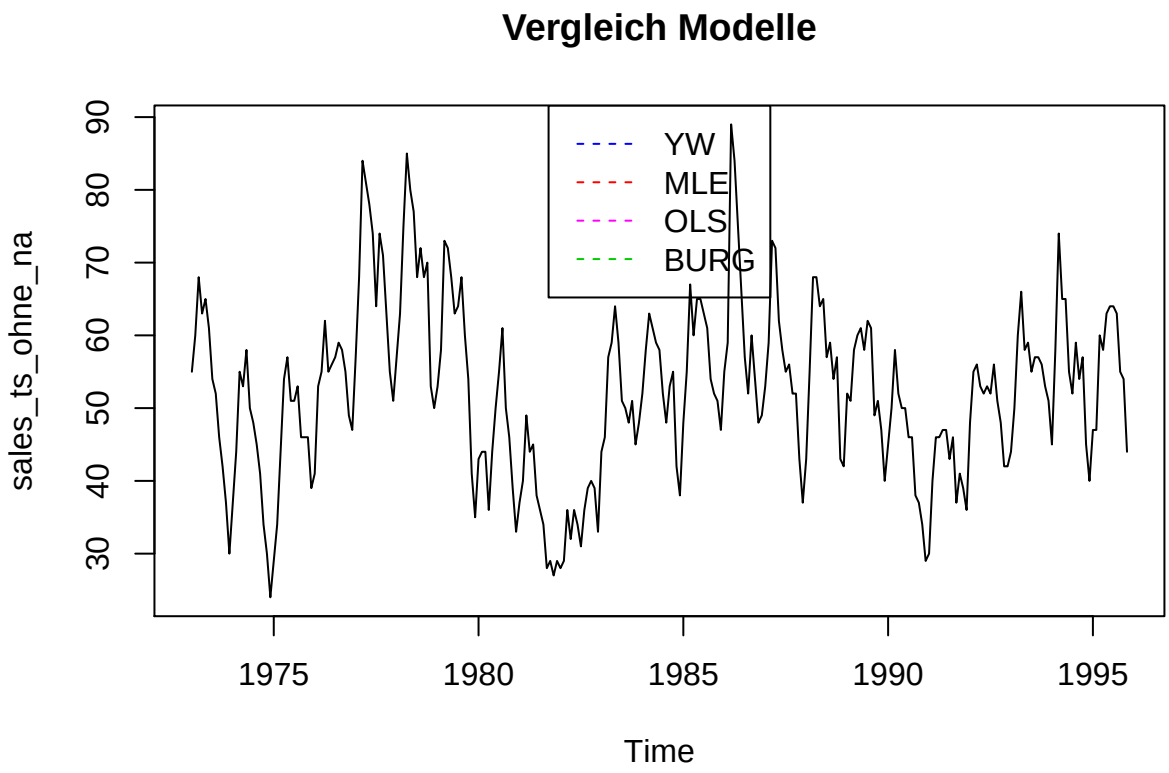
```
#Comparing Graphical order with every method
aic=mm1_mle$aic  # For plotting below.
plot(c(0:(length(aic)-1)),aic,type='h',xlab='order',ylab='aic')
lines(0:(length(mm2_yw$aic)-1), mm2_yw$aic, lty=4, col=2)
lines(0:(length(mm3_ols$aic)-1), mm3_ols$aic, lty=4, col=3)
lines(0:(length(mm4_burg$aic)-1), mm4_burg$aic, lty=4, col=4)
lines(0:(length(aic)-1),aic,lty=2)
colors <- c("green","blue","red", "black")
legend("topright", c(paste("OLS"),
                     paste("Burg"),
                     paste("YW"),
                     paste("MLE")),
       lwd = 1, cex=1, col=colors, pt.lwd = 1, lty=4)
```



Graphical comparison of the models:

```
#Comparing the graphical results
plot(sales_ts_ohne_na, main="Vergleich Modelle")
lines(fitted(mm2_yw),col="blue")
lines(fitted(mm1_mle),col="red")
lines(fitted(mm3_ols),col=6)
lines(fitted(mm4_burg),col=3)
colors <- c("blue","red",6, 3)
legend("top", c(paste("YW"),
                paste("MLE"),
                paste("OLS"),
```

```
                paste("BURG")),
       lwd = 1, cex=1, col=colors, pt.lwd = 1, lty=2)
```



3. The good of the models directly with `sarima()` with `astsa` package:

Which model is best? With AIC or BIC:

- More about AIC: https://en.wikipedia.org/wiki/Akaike_information_criterion
- More about BIC: https://en.wikipedia.org/wiki/Bayesian_information_criterion

AIC:

```
sarima(sales_ts_ohne_na,10,0,0)$AIC
```

See Figure 6.61

```
# initial  value 2.485953
# iter   2 value 2.178505
# iter   3 value 1.991777
# iter   4 value 1.956337
# iter   5 value 1.948356
# iter   6 value 1.829770
# iter   7 value 1.678623
# iter   8 value 1.665761
# iter   9 value 1.653535
# iter  10 value 1.650139
# iter  11 value 1.647349
# iter  12 value 1.646675
# iter  13 value 1.646441
```

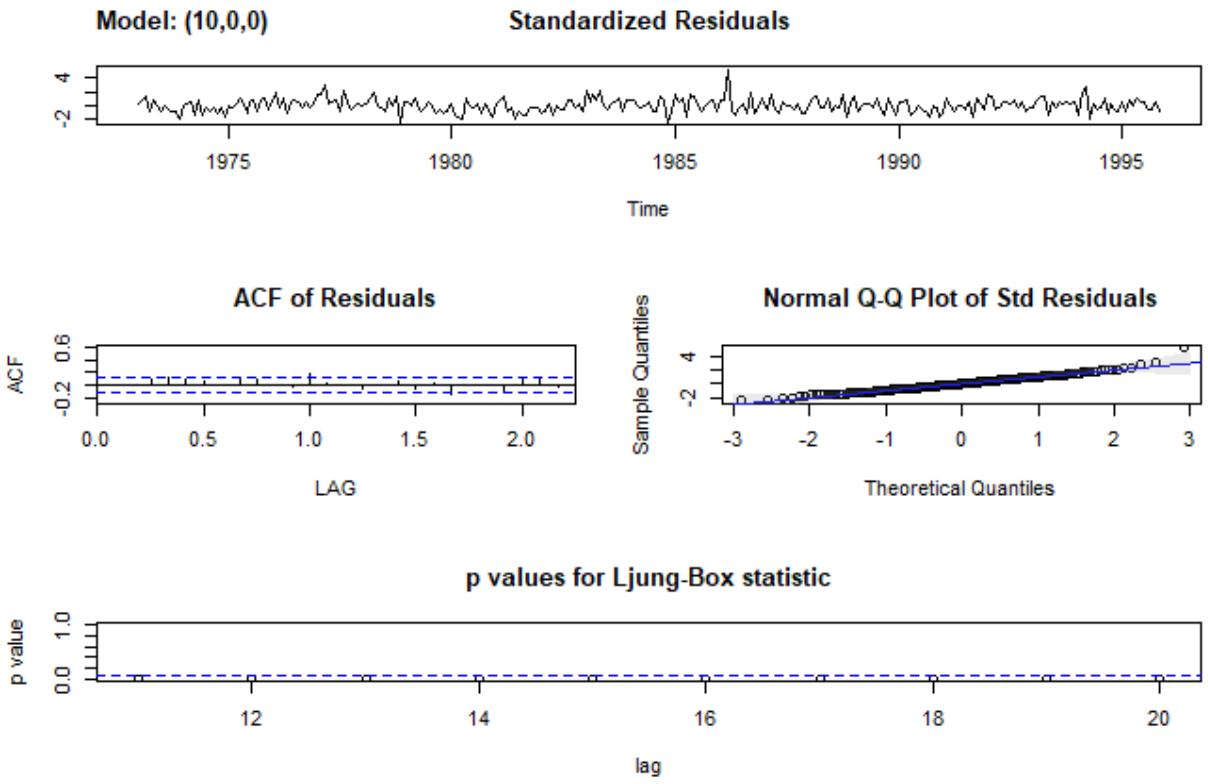

Model: (10,0,0)
Standardized Residuals
-2
4
1975
1980
1985
1990
1995
Time
ACF of Residuals
ACF
-0.2
0.6
0.0
0.5
1.0
1.5
2.0
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
-2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0
1.0
12
14
16
18
20
lag

Figure 6.59:

```
# iter  14 value 1.646403
# iter  15 value 1.646374
# iter  16 value 1.646371
# iter  17 value 1.646369
# iter  18 value 1.646369
# iter  19 value 1.646369
# iter  19 value 1.646369
# iter  19 value 1.646369
# final  value 1.646369
# converged
# initial  value 1.645266
# iter   2 value 1.645189
# iter   3 value 1.645112
# iter   4 value 1.645071
# iter   5 value 1.645054
# iter   6 value 1.645051
# iter   7 value 1.645042
# iter   8 value 1.645038
# iter   9 value 1.645036
# iter  10 value 1.645036
# iter  11 value 1.645035
# iter  12 value 1.645033
# iter  13 value 1.645030
# iter  14 value 1.645027
# iter  15 value 1.645025
# iter  16 value 1.645025
# iter  17 value 1.645025
# iter  18 value 1.645025
# iter  18 value 1.645025
# iter  18 value 1.645025
# final  value 1.645025
# converged
# [1] 4.356501
```

```
sarima(sales_ts,15,0,1)$AIC
```

See Figure 6.62

```
# initial  value 2.477969
# iter   2 value 2.072883
# iter   3 value 1.821983
# iter   4 value 1.681173
# iter   5 value 1.639634
# iter   6 value 1.613854
# iter   7 value 1.595848
# iter   8 value 1.589525
# iter   9 value 1.579773
# iter  10 value 1.577979
# iter  11 value 1.576519
# iter  12 value 1.576082
# iter  13 value 1.575352
# iter  14 value 1.573996
# iter  15 value 1.573032
# iter  16 value 1.572449
# iter  17 value 1.571974
```

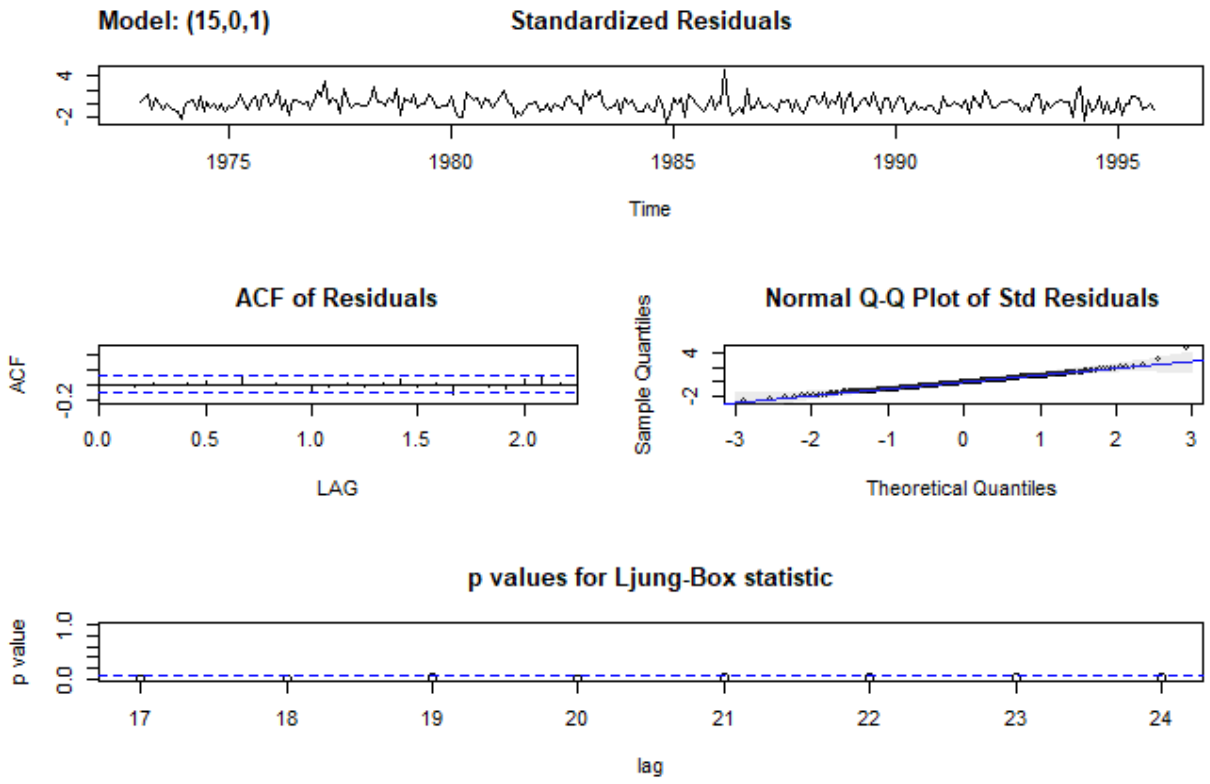

Model: (15,0,1)
Standardized Residuals
-2
4
1975
1980
1985
1990
1995
Time
ACF of Residuals
ACF
-0.2
0.0
0.5
1.0
1.5
2.0
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
-2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0
1.0
17
18
19
20
21
22
23
24
lag

Figure 6.60:

```
# iter  18 value 1.571453
# iter  19 value 1.570694
# iter  20 value 1.570000
# iter  21 value 1.569741
# iter  22 value 1.569332
# iter  23 value 1.569077
# iter  24 value 1.568539
# iter  25 value 1.568131
# iter  26 value 1.567837
# iter  27 value 1.567613
# iter  28 value 1.567547
# iter  29 value 1.567314
# iter  30 value 1.566957
# iter  31 value 1.566464
# iter  32 value 1.566286
# iter  33 value 1.566007
# iter  34 value 1.565913
# iter  35 value 1.565642
# iter  36 value 1.565329
# iter  37 value 1.565165
# iter  38 value 1.565043
# iter  39 value 1.564868
# iter  40 value 1.564748
# iter  41 value 1.564563
# iter  42 value 1.564460
# iter  43 value 1.564341
# iter  44 value 1.564246
# iter  45 value 1.564093
# iter  46 value 1.563823
# iter  47 value 1.563560
# iter  48 value 1.563101
# iter  49 value 1.562926
# iter  50 value 1.562624
# iter  51 value 1.562447
# iter  52 value 1.562319
# iter  53 value 1.562205
# iter  54 value 1.562022
# iter  55 value 1.561951
# iter  56 value 1.561723
# iter  57 value 1.561721
# iter  58 value 1.561610
# iter  59 value 1.561586
# iter  60 value 1.561526
# iter  61 value 1.561456
# iter  62 value 1.561427
# iter  63 value 1.561362
# iter  64 value 1.561224
# iter  65 value 1.561061
# iter  66 value 1.560970
# iter  67 value 1.560902
# iter  68 value 1.560786
# iter  69 value 1.560540
# iter  70 value 1.560241
```

```
# iter  71 value 1.559551
# iter  72 value 1.557250
# iter  73 value 1.556947
# iter  74 value 1.556737
# iter  75 value 1.555458
# iter  76 value 1.554503
# iter  77 value 1.554201
# iter  78 value 1.553312
# iter  79 value 1.552252
# iter  80 value 1.551811
# iter  81 value 1.551653
# iter  82 value 1.551361
# iter  83 value 1.551201
# iter  84 value 1.551148
# iter  85 value 1.551120
# iter  86 value 1.551110
# iter  87 value 1.551105
# iter  88 value 1.551102
# iter  89 value 1.551102
# iter  90 value 1.551101
# iter  91 value 1.551101
# iter  91 value 1.551101
# iter  91 value 1.551101
# final  value 1.551101
# converged
# [1] 4.192116
```

```
sarima(sales_ts_ohne_na,21,0,1)$AIC
```

See Figure 6.63

```
# initial  value 2.490162
# iter   2 value 2.065409
# iter   3 value 1.861024
# iter   4 value 1.727229
# iter   5 value 1.615131
# iter   6 value 1.564840
# iter   7 value 1.553164
# iter   8 value 1.547655
# iter   9 value 1.542343
# iter  10 value 1.540627
# iter  11 value 1.540452
# iter  12 value 1.539995
# iter  13 value 1.539881
# iter  14 value 1.539646
# iter  15 value 1.539150
# iter  16 value 1.538334
# iter  17 value 1.537251
# iter  18 value 1.536838
# iter  19 value 1.536819
# iter  20 value 1.536651
# iter  21 value 1.536637
# iter  22 value 1.536619
# iter  23 value 1.536591
# iter  24 value 1.536567
```

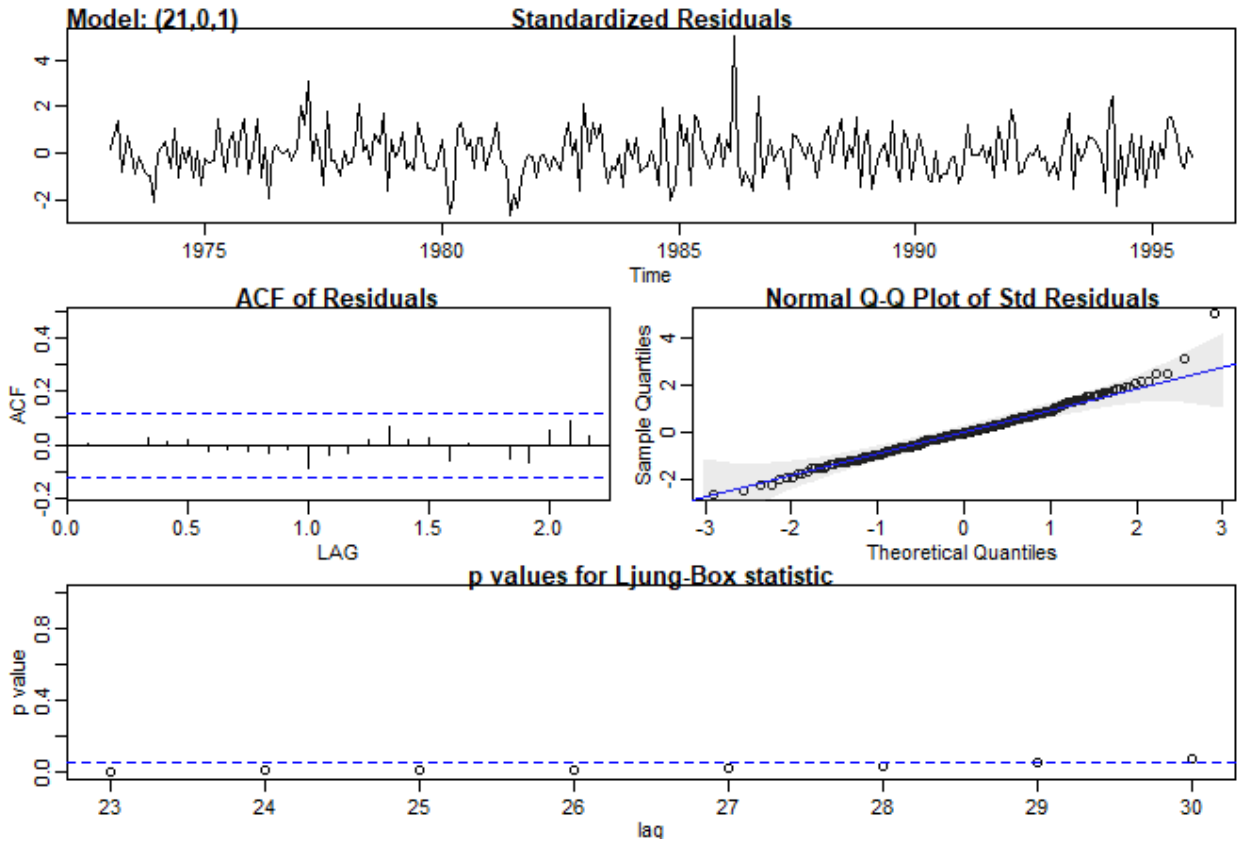

Model: (21,0,1)
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value
lag

Figure 6.61:

```
# iter  25 value 1.536548
# iter  26 value 1.536537
# iter  27 value 1.536530
# iter  28 value 1.536528
# iter  29 value 1.536527
# iter  30 value 1.536527
# iter  31 value 1.536527
# iter  31 value 1.536527
# iter  31 value 1.536527
# final  value 1.536527
# converged
# initial  value 1.536697
# iter   2 value 1.536583
# iter   3 value 1.536506
# iter   4 value 1.536406
# iter   5 value 1.536354
# iter   6 value 1.536300
# iter   7 value 1.536229
# iter   8 value 1.536208
# iter   9 value 1.536196
# iter  10 value 1.536191
# iter  11 value 1.536189
# iter  12 value 1.536188
# iter  13 value 1.536187
# iter  14 value 1.536186
# iter  15 value 1.536184
# iter  16 value 1.536183
# iter  17 value 1.536181
# iter  18 value 1.536178
# iter  19 value 1.536176
# iter  20 value 1.536174
# iter  21 value 1.536172
# iter  22 value 1.536171
# iter  23 value 1.536170
# iter  24 value 1.536168
# iter  25 value 1.536168
# iter  26 value 1.536167
# iter  27 value 1.536166
# iter  28 value 1.536166
# iter  29 value 1.536165
# iter  30 value 1.536165
# iter  30 value 1.536165
# iter  30 value 1.536165
# final  value 1.536165
# converged
# [1] 4.212936
```

BIC:

```
sarima(sales_ts_ohne_na,10,0,0)$BIC
```

See Figure 6.64

```
# initial  value 2.485953
# iter   2 value 2.178505
```

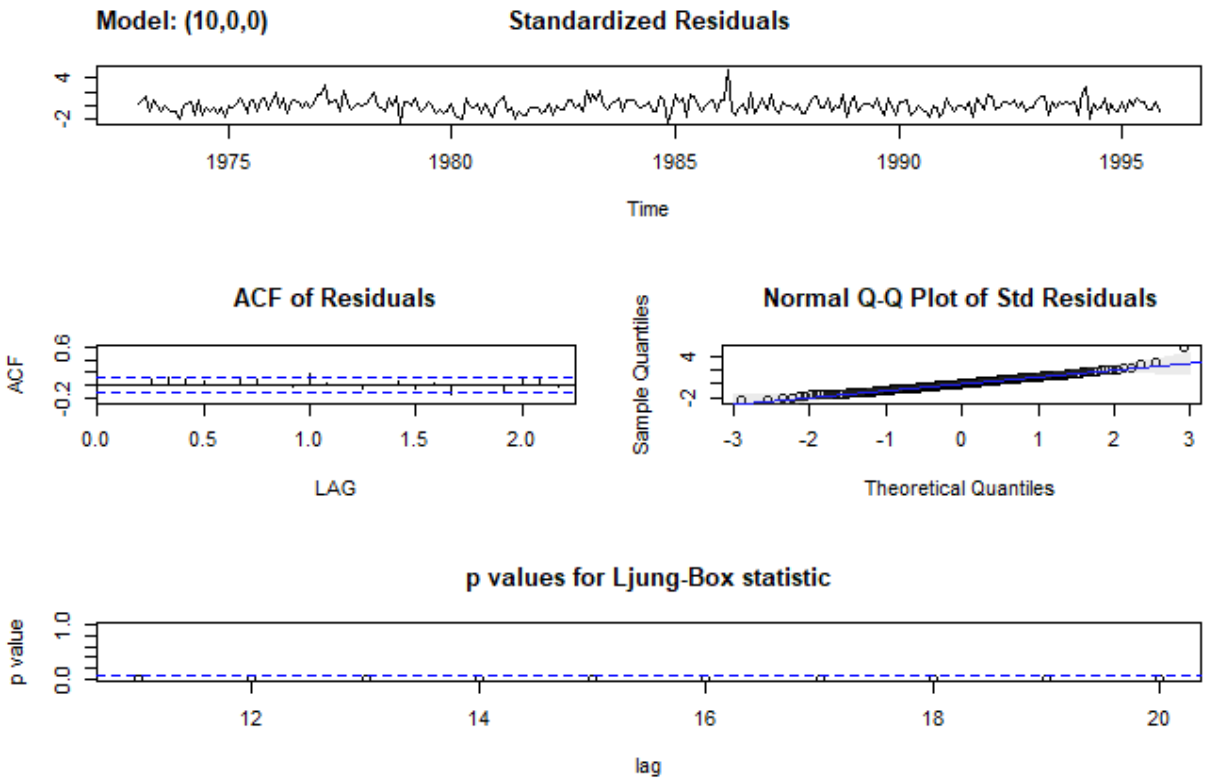

Model: (10,0,0)
Standardized Residuals
-2
4
1975
1980
1985
1990
1995
Time
ACF of Residuals
ACF
-0.2
0.6
0.0
0.5
1.0
1.5
2.0
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
-2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0
1.0
12
14
16
18
20
lag

Figure 6.62:

```
# iter   3 value 1.991777
# iter   4 value 1.956337
# iter   5 value 1.948356
# iter   6 value 1.829770
# iter   7 value 1.678623
# iter   8 value 1.665761
# iter   9 value 1.653535
# iter  10 value 1.650139
# iter  11 value 1.647349
# iter  12 value 1.646675
# iter  13 value 1.646441
# iter  14 value 1.646403
# iter  15 value 1.646374
# iter  16 value 1.646371
# iter  17 value 1.646369
# iter  18 value 1.646369
# iter  19 value 1.646369
# iter  19 value 1.646369
# iter  19 value 1.646369
# final  value 1.646369
# converged
# initial  value 1.645266
# iter   2 value 1.645189
# iter   3 value 1.645112
# iter   4 value 1.645071
# iter   5 value 1.645054
# iter   6 value 1.645051
# iter   7 value 1.645042
# iter   8 value 1.645038
# iter   9 value 1.645036
# iter  10 value 1.645036
# iter  11 value 1.645035
# iter  12 value 1.645033
# iter  13 value 1.645030
# iter  14 value 1.645027
# iter  15 value 1.645025
# iter  16 value 1.645025
# iter  17 value 1.645025
# iter  18 value 1.645025
# iter  18 value 1.645025
# iter  18 value 1.645025
# final  value 1.645025
# converged
# [1] 3.501172
```

```
sarima(sales_ts,15,0,1)$BIC
```

See Figure 6.65

```
# initial  value 2.477969
# iter   2 value 2.072883
# iter   3 value 1.821983
# iter   4 value 1.681173
# iter   5 value 1.639634
# iter   6 value 1.613854
```

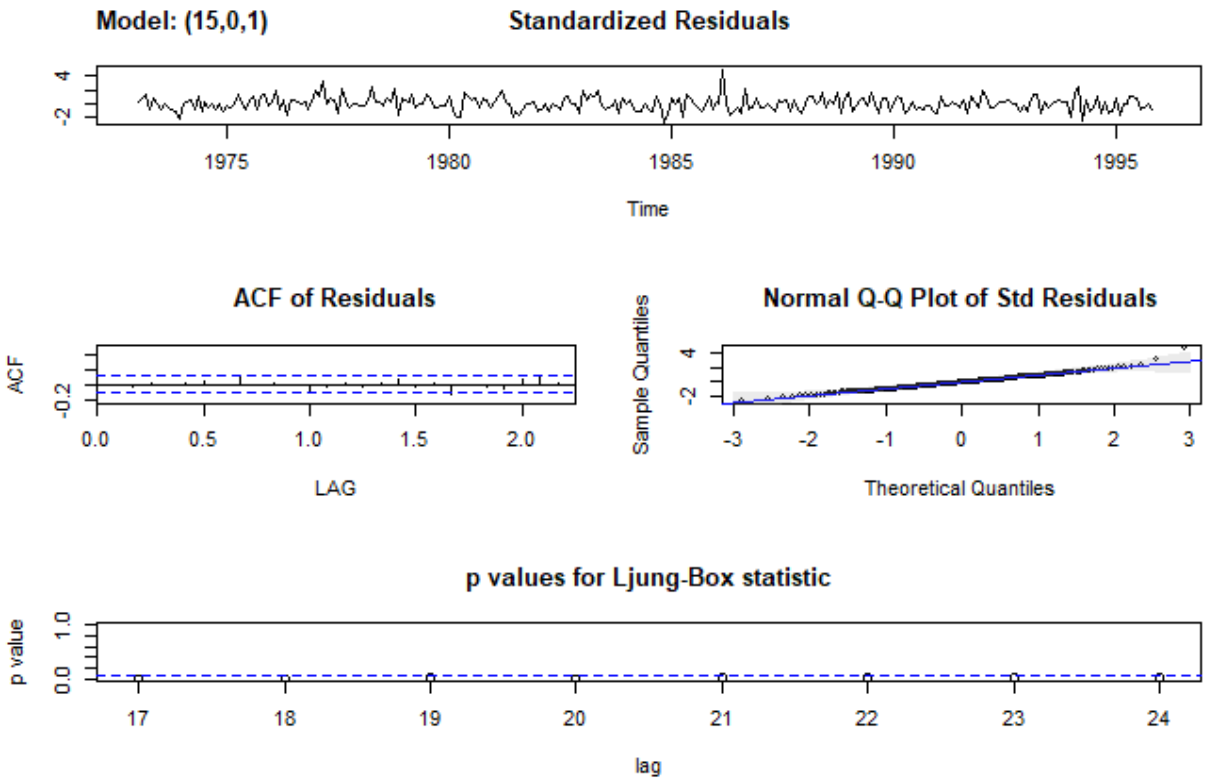

Model: (15,0,1)
Standardized Residuals
-2
4
1975
1980
1985
1990
1995
Time
ACF of Residuals
ACF
-0.2
0.0
0.5
1.0
1.5
2.0
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
-2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0
1.0
17
18
19
20
21
22
23
24
lag

Figure 6.63:

```
# iter   7 value 1.595848
# iter   8 value 1.589525
# iter   9 value 1.579773
# iter  10 value 1.577979
# iter  11 value 1.576519
# iter  12 value 1.576082
# iter  13 value 1.575352
# iter  14 value 1.573996
# iter  15 value 1.573032
# iter  16 value 1.572449
# iter  17 value 1.571974
# iter  18 value 1.571453
# iter  19 value 1.570694
# iter  20 value 1.570000
# iter  21 value 1.569741
# iter  22 value 1.569332
# iter  23 value 1.569077
# iter  24 value 1.568539
# iter  25 value 1.568131
# iter  26 value 1.567837
# iter  27 value 1.567613
# iter  28 value 1.567547
# iter  29 value 1.567314
# iter  30 value 1.566957
# iter  31 value 1.566464
# iter  32 value 1.566286
# iter  33 value 1.566007
# iter  34 value 1.565913
# iter  35 value 1.565642
# iter  36 value 1.565329
# iter  37 value 1.565165
# iter  38 value 1.565043
# iter  39 value 1.564868
# iter  40 value 1.564748
# iter  41 value 1.564563
# iter  42 value 1.564460
# iter  43 value 1.564341
# iter  44 value 1.564246
# iter  45 value 1.564093
# iter  46 value 1.563823
# iter  47 value 1.563560
# iter  48 value 1.563101
# iter  49 value 1.562926
# iter  50 value 1.562624
# iter  51 value 1.562447
# iter  52 value 1.562319
# iter  53 value 1.562205
# iter  54 value 1.562022
# iter  55 value 1.561951
# iter  56 value 1.561723
# iter  57 value 1.561721
# iter  58 value 1.561610
# iter  59 value 1.561586
```

```
# iter  60 value 1.561526
# iter  61 value 1.561456
# iter  62 value 1.561427
# iter  63 value 1.561362
# iter  64 value 1.561224
# iter  65 value 1.561061
# iter  66 value 1.560970
# iter  67 value 1.560902
# iter  68 value 1.560786
# iter  69 value 1.560540
# iter  70 value 1.560241
# iter  71 value 1.559551
# iter  72 value 1.557250
# iter  73 value 1.556947
# iter  74 value 1.556737
# iter  75 value 1.555458
# iter  76 value 1.554503
# iter  77 value 1.554201
# iter  78 value 1.553312
# iter  79 value 1.552252
# iter  80 value 1.551811
# iter  81 value 1.551653
# iter  82 value 1.551361
# iter  83 value 1.551201
# iter  84 value 1.551148
# iter  85 value 1.551120
# iter  86 value 1.551110
# iter  87 value 1.551105
# iter  88 value 1.551102
# iter  89 value 1.551102
# iter  90 value 1.551101
# iter  91 value 1.551101
# iter  91 value 1.551101
# iter  91 value 1.551101
# final  value 1.551101
# converged
# [1] 3.414529
```

```
sarima(sales_ts_ohne_na,21,0,1)$BIC
```

See Figure 6.66

```
# initial  value 2.490162
# iter   2 value 2.065409
# iter   3 value 1.861024
# iter   4 value 1.727229
# iter   5 value 1.615131
# iter   6 value 1.564840
# iter   7 value 1.553164
# iter   8 value 1.547655
# iter   9 value 1.542343
# iter  10 value 1.540627
# iter  11 value 1.540452
# iter  12 value 1.539995
# iter  13 value 1.539881
```

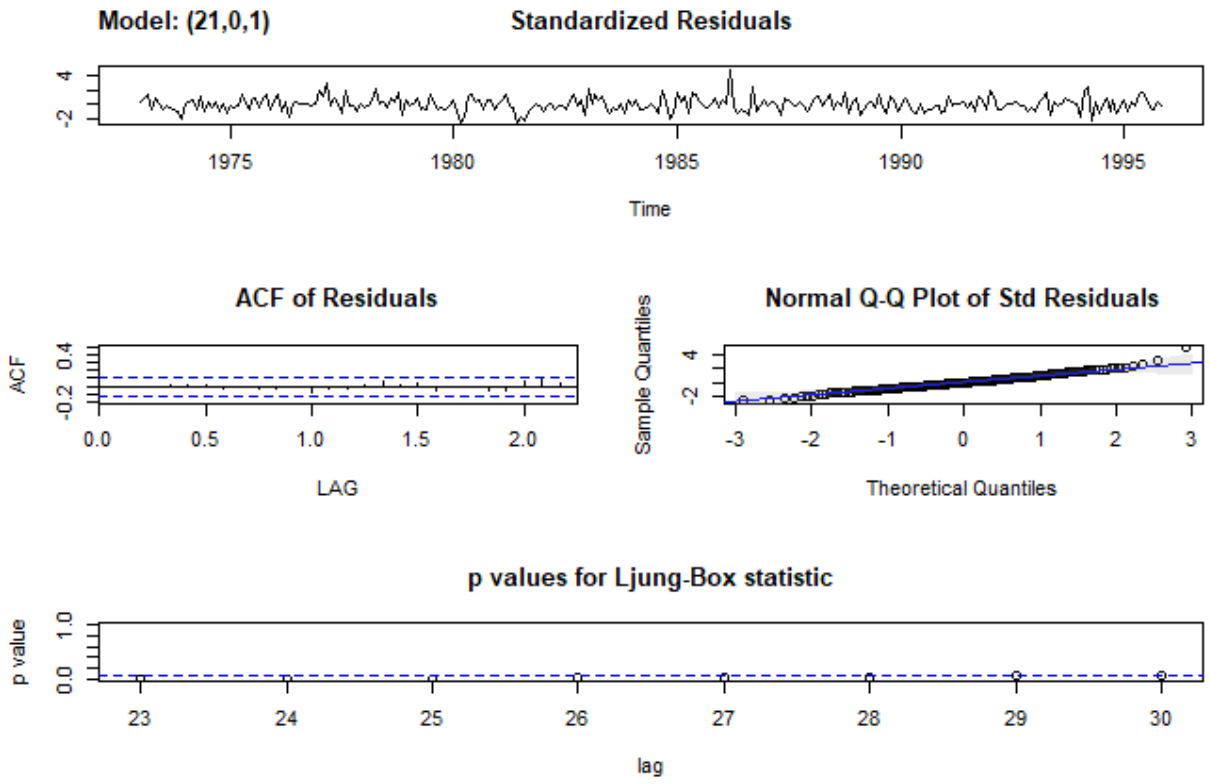

Model: (21,0,1)
Standardized Residuals
-2
4
1975
1980
1985
1990
1995
Time
ACF of Residuals
ACF
-0.2
0.4
0.0
0.5
1.0
1.5
2.0
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
-2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0
1.0
23
24
25
26
27
28
29
30
lag

Figure 6.64:

```
# iter  14 value 1.539646
# iter  15 value 1.539150
# iter  16 value 1.538334
# iter  17 value 1.537251
# iter  18 value 1.536838
# iter  19 value 1.536819
# iter  20 value 1.536651
# iter  21 value 1.536637
# iter  22 value 1.536619
# iter  23 value 1.536591
# iter  24 value 1.536567
# iter  25 value 1.536548
# iter  26 value 1.536537
# iter  27 value 1.536530
# iter  28 value 1.536528
# iter  29 value 1.536527
# iter  30 value 1.536527
# iter  31 value 1.536527
# iter  31 value 1.536527
# iter  31 value 1.536527
# final  value 1.536527
# converged
# initial  value 1.536697
# iter   2 value 1.536583
# iter   3 value 1.536506
# iter   4 value 1.536406
# iter   5 value 1.536354
# iter   6 value 1.536300
# iter   7 value 1.536229
# iter   8 value 1.536208
# iter   9 value 1.536196
# iter  10 value 1.536191
# iter  11 value 1.536189
# iter  12 value 1.536188
# iter  13 value 1.536187
# iter  14 value 1.536186
# iter  15 value 1.536184
# iter  16 value 1.536183
# iter  17 value 1.536181
# iter  18 value 1.536178
# iter  19 value 1.536176
# iter  20 value 1.536174
# iter  21 value 1.536172
# iter  22 value 1.536171
# iter  23 value 1.536170
# iter  24 value 1.536168
# iter  25 value 1.536168
# iter  26 value 1.536167
# iter  27 value 1.536166
# iter  28 value 1.536166
# iter  29 value 1.536165
# iter  30 value 1.536165
# iter  30 value 1.536165
```

```
# iter  30 value 1.536165
# final  value 1.536165
# converged
# [1] 3.51543
```

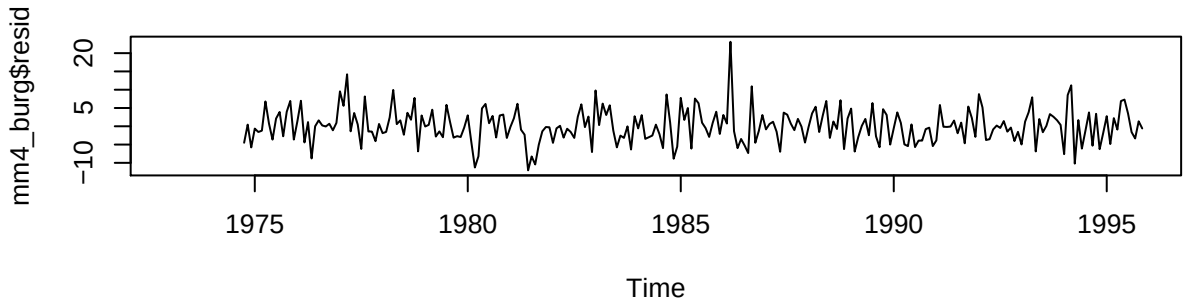
The WINNER:Grade 21

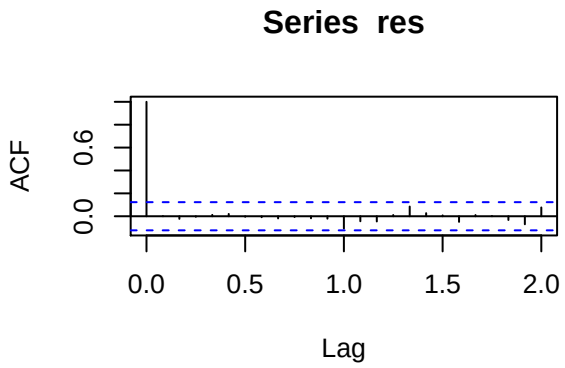
4. Residual test for best model (Order 21):

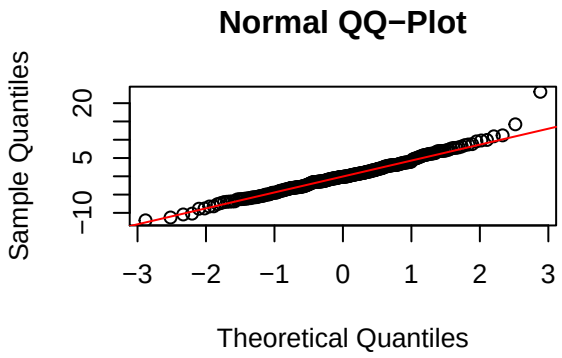
```
#Residual with mm4_burg oder mm3_ols
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
plot(mm4_burg$resid)
res=na.omit(mm4_burg$resid)
acf(res)
Box.test(res, lag=20, type="Ljung-Box", fitdf=2)
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 7.7008, df = 18, p-value = 0.9828
```

```
qqnorm(res, main="Normal QQ-Plot")  # qq-plots
qqline(res, col=2)
```

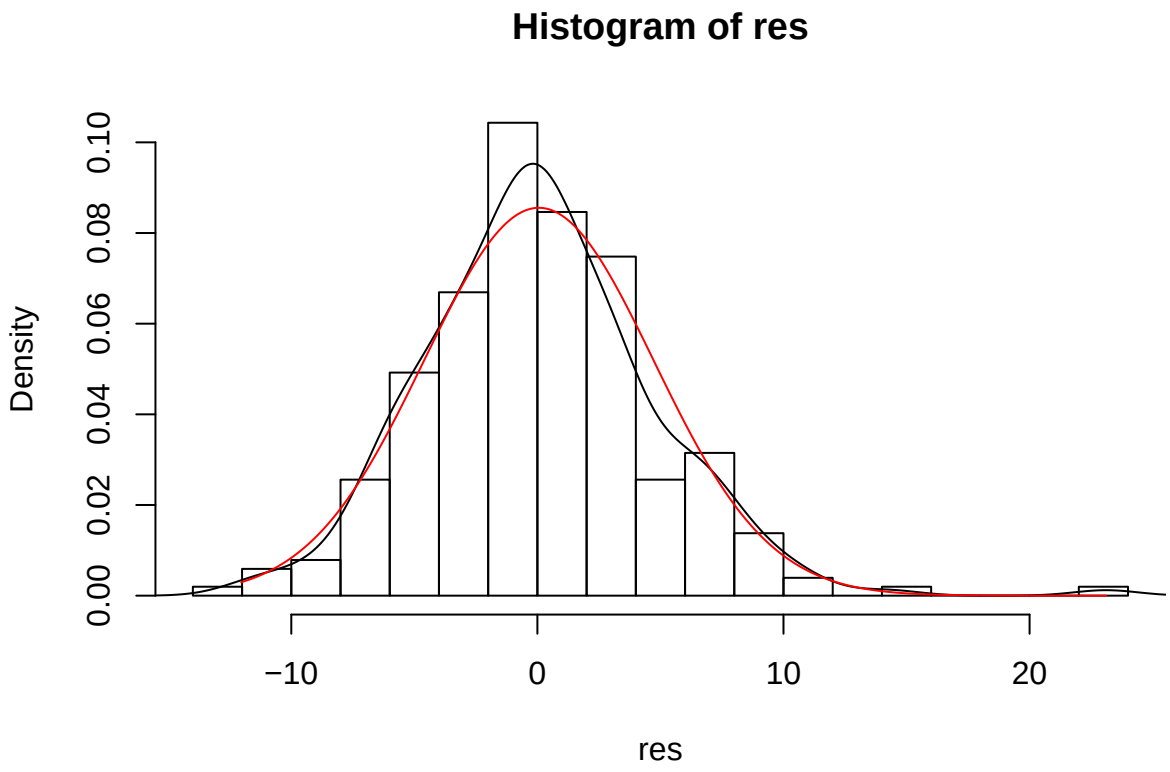







```
par(mfrow=c(1,1))
hist(res, prob=TRUE,16)   # histogram
lines(density(res))
```

```
dn=dnorm(x=seq(min(res),max(res),length.out=500), mean(res), sd(res))
lines(x=seq(min(res),max(res),length.out=500), dn, col=2)
```




```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97839, p-value = 0.0006571
```

ARIMA Modelle (`auto.arima()`):

`auto.arima()` from `forecast` package recognizes the models automatically.

```
fit_auto_arima<-auto.arima(sales_ts_ohne_na)
```

```
# Series: sales_ts_ohne_na
# ARIMA(1,0,0)(1,1,0)[12] with drift
#
# Coefficients:
#          ar1     sar1    drift
#       0.8867  -0.4320  -0.0228
# s.e.  0.0294   0.0569   0.1642
#
# sigma^2 estimated as 27.92:  log likelihood=-811.38
# AIC=1630.76   AICc=1630.92   BIC=1645.05
```

6.5.5 Forecasts

With `STL()` function of `forecast` package:

```
# ########## We need forecast package####################
#
# real=ts(c(1973, 275), frequency = 12,
#         start = c(1995,1))
# par(mfrow=c(1,1))
# dec.pass=stl(sales_ts_ohne_na, "per")
# dec.pass
# plot(dec.pass)
# #plot daten und modell
# plot(sales_ts_ohne_na, main="Passagierzahlen: saisonale Zerlegung mit stl", ylab="Passagierzahlen in
# lines((dec.pass$time.series[,2]), col=4)
# lines((dec.pass$time.series[,2]+dec.pass$time.series[,1]), col=2)
#
# plot(forecast(dec.pass), main="Forecast stl")
#
#
# plot(sales_ts_ohne_na, main="Monatlichen Sales: Vorhersage")
# lines((forecast(dec.pass)$mean), col="green", lwd=2)
# lines((forecast(dec.pass)$lower[,2]), col="blue", lwd=2)
# lines((forecast(dec.pass)$upper[,2]), col="blue", lwd=2)
# #lines(real, col=6, lwd=2)
```

```
### real=ts(c(1973, 275), frequency = 12,
###         start = c(1995,1))
### par(mfrow=c(1,1))
### dec.pass=stl(sales_ts_ohne_na, "per")
### dec.pass



# Call:
#  stl(x = sales_ts_ohne_na, s.window = "per")
#
# Components
#            seasonal    trend    remainder
# Jan 1973  -7.052492 61.82257   0.22992302
# Feb 1973  -1.230382 60.22258   1.00779851
# Mar 1973   9.069991 58.62260   0.30741135
# Apr 1973   7.967740 57.06243  -2.03016662
# May 1973   7.343751 55.50225   2.15399468
# Jun 1973   4.752665 54.00072   2.24661366
# Jul 1973   1.552883 52.49919  -0.05207051
# Aug 1973   2.915340 51.05776  -1.97309725
# Sep 1973  -1.374374 49.61633  -2.24195230
# Oct 1973  -2.458432 48.47375  -4.01531472
# Nov 1973  -9.064229 47.33117  -1.26693669
# Dec 1973 -12.422461 46.71048  -4.28801925
# Jan 1974  -7.052492 46.08979  -2.03730284
# Feb 1974  -1.230382 45.68398  -0.45359851
# Mar 1974   9.069991 45.27817   0.65184318
# Apr 1974   7.967740 44.83067   0.20159186
```

```
# May 1974   7.343751 44.38317   6.27307982
# Jun 1974   4.752665 43.73419   1.51314832
# Jul 1974   1.552883 43.08520   3.36191367
# Aug 1974   2.915340 42.27816  -0.19349857
# Sep 1974  -1.374374 41.47111   0.90326087
# Oct 1974  -2.458432 40.95995  -4.50151810
# Nov 1974  -9.064229 40.44879  -1.38455662
# Dec 1974 -12.422461 40.52756  -4.10509649
# Jan 1975  -7.052492 40.60633  -4.55383740
# Feb 1975  -1.230382 41.25403  -6.02364619
# Mar 1975   9.069991 41.90173  -6.97171763
# Apr 1975   7.967740 42.99659   3.03567358
# May 1975   7.343751 44.09145   5.56480408
# Jun 1975   4.752665 45.38910   0.85823284
# Jul 1975   1.552883 46.68676   2.76035846
# Aug 1975   2.915340 47.76362   2.32103732
# Sep 1975  -1.374374 48.84049  -1.46611212
# Oct 1975  -2.458432 49.39289  -0.93446019
# Nov 1975  -9.064229 49.94530   5.11893219
# Dec 1975 -12.422461 50.26758   1.15487730
# Jan 1976  -7.052492 50.58987  -2.53737863
# Feb 1976  -1.230382 51.09301   3.13737254
# Mar 1976   9.069991 51.59615  -5.66613893
# Apr 1976   7.967740 52.24755   1.78470749
# May 1976   7.343751 52.89896  -5.24270682
# Jun 1976   4.752665 53.78671  -2.53937448
# Jul 1976   1.552883 54.67446   0.77265471
# Aug 1976   2.915340 56.14859  -0.06392721
# Sep 1976  -1.374374 57.62271   1.75166257
# Oct 1976  -2.458432 59.43175  -1.97331844
# Nov 1976  -9.064229 61.24079  -3.17655900
# Dec 1976 -12.422461 62.77446  -3.35199618
# Jan 1977  -7.052492 64.30813  -0.25563440
# Feb 1977  -1.230382 65.44889   3.78148986
# Mar 1977   9.069991 66.58966   8.34035148
# Apr 1977   7.967740 67.32774   5.70452219
# May 1977   7.343751 68.06582   2.59043218
# Jun 1977   4.752665 68.18920   1.05813960
# Jul 1977   1.552883 68.31257  -5.86545612
# Aug 1977   2.915340 67.95156   3.13310168
# Sep 1977  -1.374374 67.59054   4.78383117
# Oct 1977  -2.458432 67.45458  -1.99615263
# Nov 1977  -9.064229 67.31863  -3.25439598
# Dec 1977 -12.422461 67.50216  -4.07970363
# Jan 1978  -7.052492 67.68570  -3.63321231
# Feb 1978  -1.230382 67.90944  -3.67905629
# Mar 1978   9.069991 68.13317  -2.20316292
# Apr 1978   7.967740 68.31711   8.71514611
# May 1978   7.343751 68.50105   4.15519441
# Jun 1978   4.752665 68.43810   3.80923053
# Jul 1978   1.552883 68.37515  -1.92803649
# Aug 1978   2.915340 67.82881   1.25584535
# Sep 1978  -1.374374 67.28248   2.09189888
```

```
# Oct 1978  -2.458432 66.41100   6.04742871
# Nov 1978  -9.064229 65.53953  -3.47530102
# Dec 1978 -12.422461 64.69248  -2.27002103
# Jan 1979  -7.052492 63.84543  -3.79294208
# Feb 1979  -1.230382 63.18734  -3.95696255
# Mar 1979   9.069991 62.52925   1.40075433
# Apr 1979   7.967740 61.77447   2.25778892
# May 1979   7.343751 61.01969  -0.36343721
# Jun 1979   4.752665 60.03600  -1.78866950
# Jul 1979   1.552883 59.05232   3.39479508
# Aug 1979   2.915340 57.45553   7.62912815
# Sep 1979  -1.374374 55.85874   5.51563292
# Oct 1979  -2.458432 53.66388   2.79454913
# Nov 1979  -9.064229 51.46902  -1.40479421
# Dec 1979 -12.422461 49.63182  -2.20935852
# Jan 1980  -7.052492 47.79462   2.25787613
# Feb 1980  -1.230382 46.81209  -1.58171047
# Mar 1980   9.069991 45.82957 -10.89955971
# Apr 1980   7.967740 45.52741 -17.49515297
# May 1980   7.343751 45.22526  -8.56900695
# Jun 1980   4.752665 45.30657  -0.05923179
# Jul 1980   1.552883 45.38788   8.05924023
# Aug 1980   2.915340 45.63098  12.45367875
# Sep 1980  -1.374374 45.87409   5.50028896
# Oct 1980  -2.458432 45.81279   2.64563820
# Nov 1980  -9.064229 45.75150   2.31272788
# Dec 1980 -12.422461 44.67895   0.74351401
# Jan 1981  -7.052492 43.60639   0.44609911
# Feb 1981  -1.230382 41.86292  -0.63254021
# Mar 1981   9.069991 40.11945  -0.18944217
# Apr 1981   7.967740 38.66696  -2.63470542
# May 1981   7.343751 37.21448   0.44177061
# Jun 1981   4.752665 36.32708  -3.07974602
# Jul 1981   1.552883 35.43968  -0.99256579
# Aug 1981   2.915340 34.73098  -3.64631921
# Sep 1981  -1.374374 34.02228  -4.64790094
# Oct 1981  -2.458432 33.31112  -1.85268400
# Nov 1981  -9.064229 32.59996   3.46427338
# Dec 1981 -12.422461 32.11584   9.30662189
# Jan 1982  -7.052492 31.63172   3.42076936
# Feb 1982  -1.230382 31.63684  -1.40645631
# Mar 1982   9.069991 31.64195  -4.71194463
# Apr 1982   7.967740 32.17310  -8.14083575
# May 1982   7.343751 32.70424  -4.04798760
# Jun 1982   4.752665 33.77989  -4.53255552
# Jul 1982   1.552883 34.85554  -5.40842659
# Aug 1982   2.915340 36.57721  -3.49255085
# Sep 1982  -1.374374 38.29888   2.07549658
# Oct 1982  -2.458432 40.44169   2.01673668
# Nov 1982  -9.064229 42.58451   5.47971723
# Dec 1982 -12.422461 44.52271   0.89975396
# Jan 1983  -7.052492 46.46090   4.59158966
# Feb 1983  -1.230382 47.72738  -0.49700185
```

```
# Mar 1983   9.069991 48.99387  -1.06385600
# Apr 1983   7.967740 49.74380   1.28846446
# May 1983   7.343751 50.49373   6.16252419
# Jun 1983   4.752665 51.24299   3.00434902
# Jul 1983   1.552883 51.99225  -2.54512929
# Aug 1983   2.915340 52.67312  -5.58846331
# Sep 1983  -1.374374 53.35400  -3.97962563
# Oct 1983  -2.458432 53.68724  -0.22880872
# Nov 1983  -9.064229 54.02048   0.04374863
# Dec 1983 -12.422461 54.08807   6.33439123
# Jan 1984  -7.052492 54.15566   4.89683279
# Feb 1984  -1.230382 54.15654   5.07383944
# Mar 1984   9.069991 54.15743  -0.22741656
# Apr 1984   7.967740 53.94335  -0.91108638
# May 1984   7.343751 53.72927  -2.07301693
# Jun 1984   4.752665 53.35159  -0.10425199
# Jul 1984   1.552883 52.97391  -2.52679020
# Aug 1984   2.915340 52.84330  -7.75864272
# Sep 1984  -1.374374 52.71270   1.66167645
# Oct 1984  -2.458432 52.99779   4.46064056
# Nov 1984  -9.064229 53.28288  -2.21865489
# Dec 1984 -12.422461 53.89935  -3.47688432
# Jan 1985  -7.052492 54.51581   0.53668521
# Feb 1985  -1.230382 55.13482   1.09556329
# Mar 1985   9.069991 55.75383   2.17617872
# Apr 1985   7.967740 56.18577  -4.15350757
# May 1985   7.343751 56.61770   1.03854542
# Jun 1985   4.752665 57.10283   3.14450950
# Jul 1985   1.552883 57.58795   3.85917043
# Aug 1985   2.915340 58.43915  -0.35449122
# Sep 1985  -1.374374 59.29036  -3.91598117
# Oct 1985  -2.458432 60.42009  -5.96166073
# Nov 1985  -9.064229 61.54983  -1.48559984
# Dec 1985 -12.422461 62.19821  -2.77574578
# Jan 1986  -7.052492 62.84658  -0.79409276
# Feb 1986  -1.230382 62.92917  -2.69879287
# Mar 1986   9.069991 63.01176  16.91824438
# Apr 1986   7.967740 62.88496  13.14729507
# May 1986   7.343751 62.75816   4.89808503
# Jun 1986   4.752665 62.35540  -1.10806094
# Jul 1986   1.552883 61.95263  -6.50551005
# Aug 1986   2.915340 61.26809 -12.18343198
# Sep 1986  -1.374374 60.58356   0.79081778
# Oct 1986  -2.458432 59.90096  -3.44252369
# Nov 1986  -9.064229 59.21835  -2.15412471
# Dec 1986 -12.422461 58.90597   2.51649340
# Jan 1987  -7.052492 58.59358   1.45891048
# Feb 1987  -1.230382 58.36275   1.86763567
# Mar 1987   9.069991 58.13191   5.79809822
# Apr 1987   7.967740 57.62558   6.40668384
# May 1987   7.343751 57.11924  -2.46299126
# Jun 1987   4.752665 56.35507  -3.10773342
# Jul 1987   1.552883 55.59090  -2.14377872
```

```
# Aug 1987   2.915340 54.96140  -1.87673866
# Sep 1987  -1.374374 54.33190  -0.95752691
# Oct 1987  -2.458432 54.20957   0.24886271
# Nov 1987  -9.064229 54.08724  -2.02300724
# Dec 1987 -12.422461 54.40582  -4.98335841
# Jan 1988  -7.052492 54.72440  -4.67191061
# Feb 1988  -1.230382 55.09893   1.13145559
# Mar 1988   9.069991 55.47345   3.45655914
# Apr 1988   7.967740 55.80992   4.22234222
# May 1988   7.343751 56.14638   0.50986457
# Jun 1988   4.752665 56.36676   3.88057176
# Jul 1988   1.552883 56.58714  -1.14002420
# Aug 1988   2.915340 56.37484  -0.29018081
# Sep 1988  -1.374374 56.16254  -0.78816573
# Oct 1988  -2.458432 55.65706   3.80136780
# Nov 1988  -9.064229 55.15159  -3.08735824
# Dec 1988 -12.422461 54.85661  -0.43414877
# Jan 1989  -7.052492 54.56163   4.49085967
# Feb 1989  -1.230382 54.46603  -2.23564647
# Mar 1989   9.069991 54.37042  -5.44041525
# Apr 1989   7.967740 54.27968  -2.24741646
# May 1989   7.343751 54.18893  -0.53267839
# Jun 1989   4.752665 54.09677  -0.84943702
# Jul 1989   1.552883 54.00462   6.44250120
# Aug 1989   2.915340 53.81896   4.26570373
# Sep 1989  -1.374374 53.63330  -3.25892204
# Oct 1989  -2.458432 53.07898   0.37945619
# Nov 1989  -9.064229 52.52465   3.53957488
# Dec 1989 -12.422461 51.54381   0.87864696
# Jan 1990  -7.052492 50.56297   1.48951801
# Feb 1990  -1.230382 49.39765   1.83273208
# Mar 1990   9.069991 48.23233   0.69768350
# Apr 1990   7.967740 47.09863  -3.06637382
# May 1990   7.343751 45.96494  -3.30869187
# Jun 1990   4.752665 44.88695   0.36038472
# Jul 1990   1.552883 43.80896   0.63815816
# Aug 1990   2.915340 42.94279   0.14187147
# Sep 1990  -1.374374 42.07662  -2.70224354
# Oct 1990  -2.458432 41.48288  -2.02444702
# Nov 1990  -9.064229 40.88914   2.17508995
# Dec 1990 -12.422461 40.59162   0.83084111
# Jan 1991  -7.052492 40.29410  -3.24160876
# Feb 1991  -1.230382 40.22752   1.00285985
# Mar 1991   9.069991 40.16094  -3.23093419
# Apr 1991   7.967740 40.41417  -2.38191515
# May 1991   7.343751 40.66741  -1.01115683
# Jun 1991   4.752665 41.41709   0.83024231
# Jul 1991   1.552883 42.16678  -0.71966169
# Aug 1991   2.915340 43.25418  -0.16951589
# Sep 1991  -1.374374 44.34157  -5.96719840
# Oct 1991  -2.458432 45.20791  -1.74947505
# Nov 1991  -9.064229 46.07424   1.98998874
# Dec 1991 -12.422461 46.74221   1.68025491
```

```
# Jan 1992  -7.052492 47.41017   7.64232004
# Feb 1992  -1.230382 48.10287   8.12751590
# Mar 1992   9.069991 48.79556  -1.86555089
# Apr 1992   7.967740 49.29445  -4.26219411
# May 1992   7.343751 49.79335  -5.13709804
# Jun 1992   4.752665 49.96677  -1.71943574
# Jul 1992   1.552883 50.14019   0.30692343
# Aug 1992   2.915340 50.35278   2.73188191
# Sep 1992  -1.374374 50.56536   1.80901208
# Oct 1992  -2.458432 51.04985  -0.59142160
# Nov 1992  -9.064229 51.53434  -0.47011483
# Dec 1992 -12.422461 51.94083   2.48162756
# Jan 1993  -7.052492 52.34732  -1.29483108
# Feb 1993  -1.230382 52.67858  -1.44820132
# Mar 1993   9.069991 53.00984  -2.07983421
# Apr 1993   7.967740 53.57410   4.45816441
# May 1993   7.343751 54.13835  -3.48209769
# Jun 1993   4.752665 54.80254  -0.55520124
# Jul 1993   1.552883 55.46673  -2.01960793
# Aug 1993   2.915340 56.11835  -2.03368553
# Sep 1993  -1.374374 56.76997   1.60440857
# Oct 1993  -2.458432 57.24291   1.21551945
# Nov 1993  -9.064229 57.71586   4.34837078
# Dec 1993 -12.422461 57.80625   5.61621212
# Jan 1994  -7.052492 57.89664  -5.84414758
# Feb 1994  -1.230382 57.66865   1.56173466
# Mar 1994   9.069991 57.44065   7.48935426
# Apr 1994   7.967740 57.06520  -0.03294373
# May 1994   7.343751 56.68975   0.96649756
# Jun 1994   4.752665 56.19324  -5.94590874
# Jul 1994   1.552883 55.69674  -5.24961819
# Aug 1994   2.915340 55.08499   0.99966537
# Sep 1994  -1.374374 54.47325   0.90112062
# Oct 1994  -2.458432 54.05382   5.40460953
# Nov 1994  -9.064229 53.63439   0.42983887
# Dec 1994 -12.422461 53.80952  -1.38705781
# Jan 1995  -7.052492 53.98465   0.06784448
# Feb 1995  -1.230382 54.40554  -6.17515534
# Mar 1995   9.069991 54.82643  -3.89641782
# Apr 1995   7.967740 55.09259  -5.06033019
# May 1995   7.343751 55.35875   0.29749670
# Jun 1995   4.752665 55.64907   3.59826236
# Jul 1995   1.552883 55.93939   6.50772488
# Aug 1995   2.915340 56.30171   3.78295467
# Sep 1995  -1.374374 56.66402  -0.28964385
# Oct 1995  -2.458432 57.05055  -0.59211603
# Nov 1995  -9.064229 57.43708  -4.37284777
```

```
plot(dec.pass)
```

See Figure 6.67

```
plot(sales_ts_ohne_na, main="Season extraction with stl", ylab="passenger numbers in
1000") lines((dec.pass$time.series[,2]), col=4) lines((dec.pass$time.series[,2]+dec.pass$time.series[,1])
col=2)
```

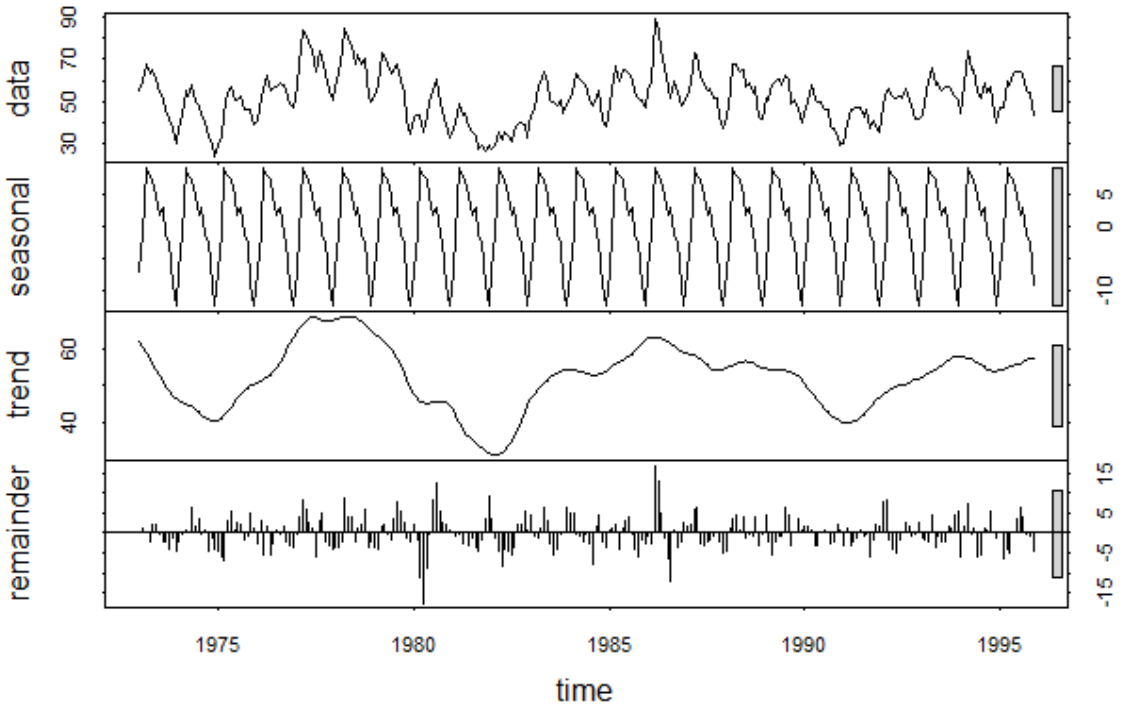

data
seasonal
trend
remainder
30
50
70
90
40
60
-10
0
5
-15
-5
5
15
1975
1980
1985
1990
1995
time

Figure 6.65:

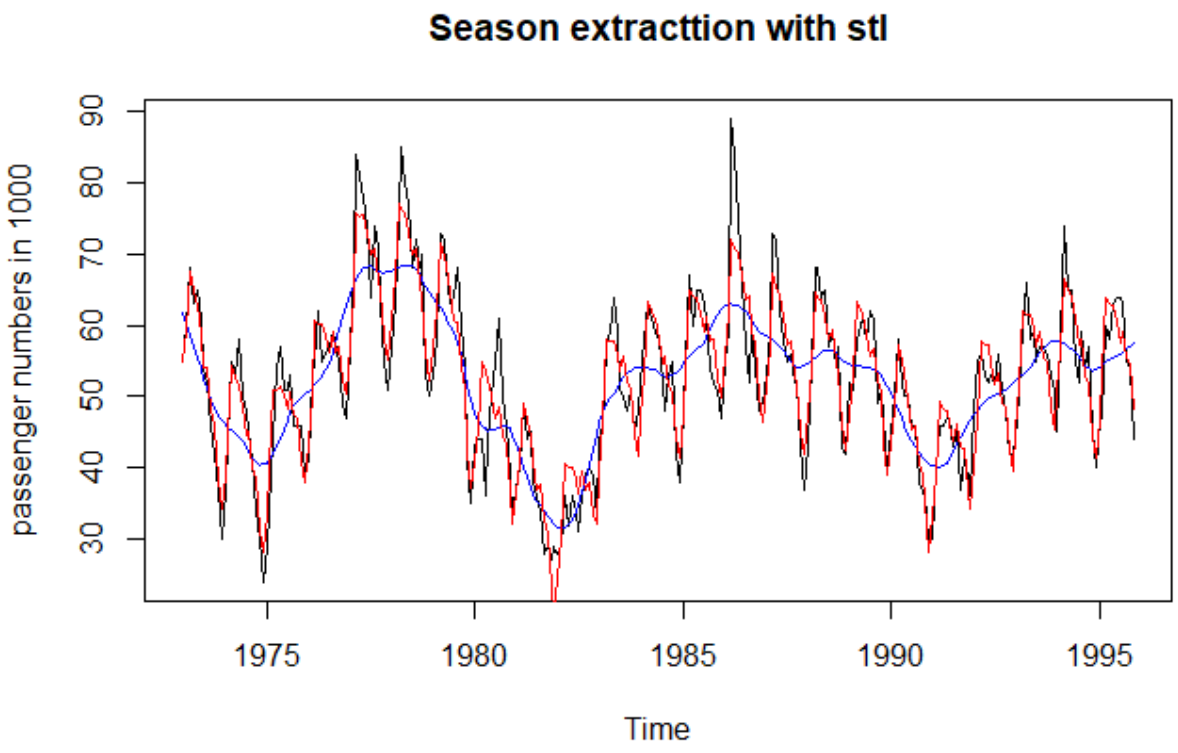


Figure 6.66:

See Figure 6.68

```
plot(forecast(dec.pass), main="Forecast stl")
```

See Figure 6.69

```
plot(sales_ts_ohne_na, main="Monatly Sales: Prediction") lines((forecast(dec.pass)$mean),
col="green", lwd=2) lines((forecast(dec.pass)$lower[,2]), col="blue", lwd=2) lines((forecast(dec.pass)$up
col="blue", lwd=2)
```

See Figure 6.70

With sarima.for() from astsa package:

```
# sarima.for(xdata, n.ahead, p, d, q, P = 0, D = 0, Q = 0, S = -1,
#             tol = sqrt(.Machine$double.eps), no.constant = FALSE,
#             plot.all=FALSE, xreg = NULL, newxreg = NULL)
```

```
sarima.for(sales_ts_ohne_na,5,1,0,0)
```

See Figure 6.71

6.5.6 Compare the models AR and Auto Arima Model

AR:

```
summary(arima(sales_ts_ohne_na,order=c(21,0,0)))
```

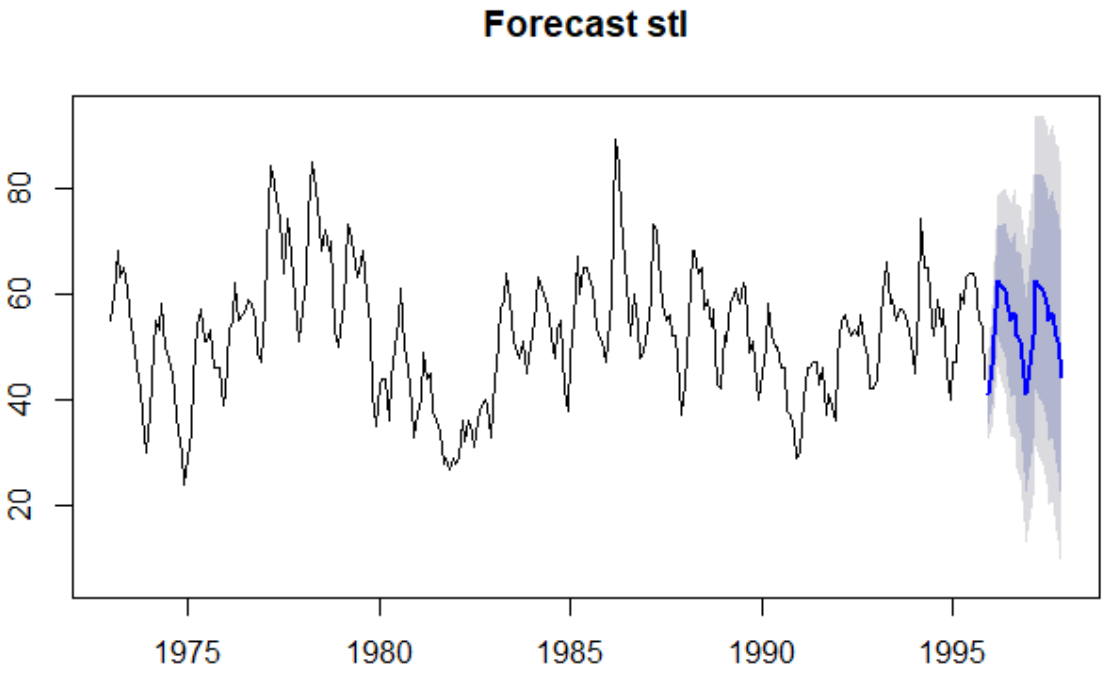


Figure 6.67:

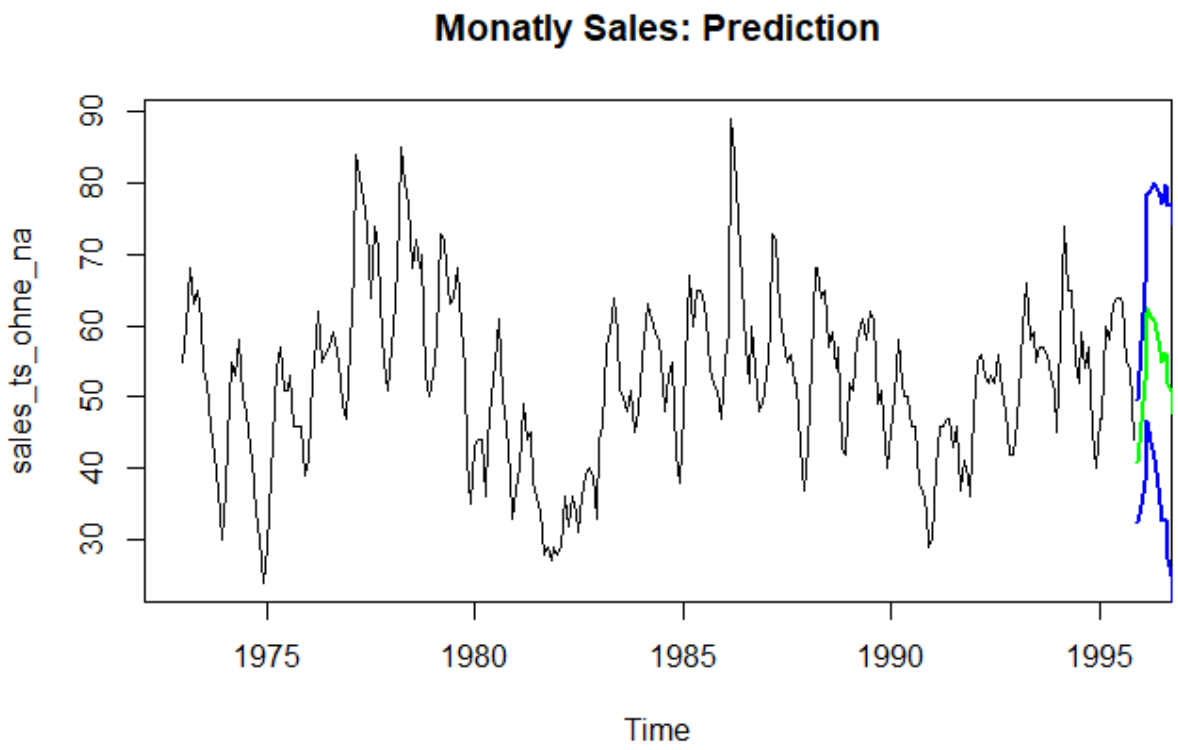

Monatly Sales: Prediction
sales_ts_ohne_na
30
40
50
60
70
80
90
1975
1980
1985
1990
1995
Time

Figure 6.68:

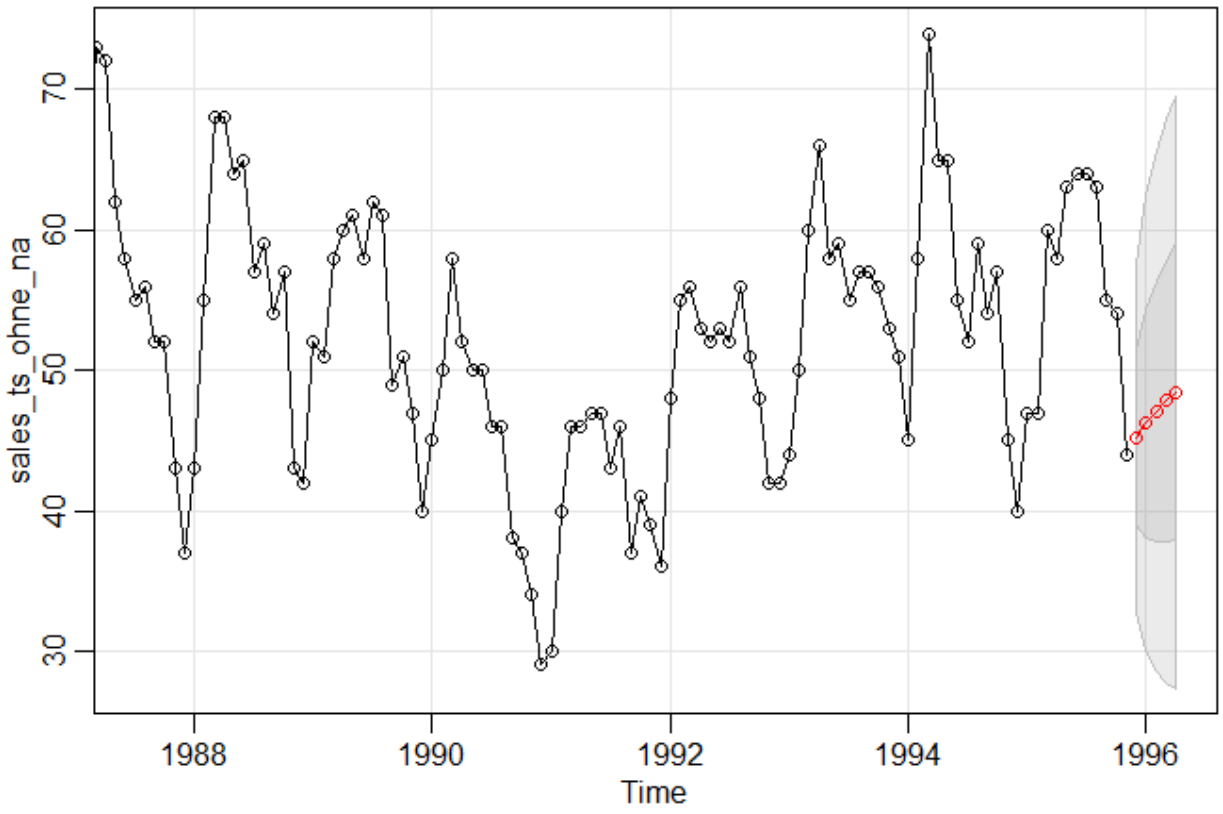


Figure 6.69:

```
# Call:
# arima(x = sales_ts_ohne_na, order = c(21, 0, 0))
#
# Coefficients:
#          ar1      ar2      ar3      ar4     ar5      ar6     ar7      ar8      ar9    ar10     ar11
#       0.8732  -0.0292  -0.0822  -0.0605  0.1595  -0.0091  0.0212  -0.0602  -0.0313  0.2014  -0.0372
# s.e.  0.0595   0.0782   0.0779   0.0775  0.0772   0.0777  0.0773   0.0773   0.0771  0.0754   0.0763
#         ar12     ar13    ar14     ar15    ar16    ar17     ar18    ar19     ar20    ar21  intercept
#       0.2892  -0.1859  0.0155  -0.1311  0.0029  0.0658  -0.1585  0.1082  -0.2335  0.1507    52.6856
# s.e.  0.0755   0.0775  0.0784   0.0784  0.0790  0.0790   0.0790  0.0796   0.0798  0.0611     2.1056
#
# sigma^2 estimated as 21.02:  log likelihood = -812.66,  aic = 1671.31
#
# Training set error measures:
#                       ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
# Training set -0.03343723 4.585222 3.501625 -0.924734 6.979855 0.6907452 0.002623198
```

Auto ARIMA:

```
summary(fit_auto_arima)
```

```
# Series: sales_ts_ohne_na
# ARIMA(1,0,0)(1,1,0)[12] with drift
#
# Coefficients:
#          ar1     sar1    drift
#       0.8867  -0.4320  -0.0228
# s.e.  0.0294   0.0569   0.1642
#
# sigma^2 estimated as 27.92:  log likelihood=-811.38
# AIC=1630.76   AICc=1630.92   BIC=1645.05
#
# Training set error measures:
#                       ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
# Training set 0.06977783 5.137588 3.982385 -0.4155256 7.894716 0.4835491 -0.03677202
```

Conclusion: Which model is the best?

- ME: Mean Error
- RMSE: root mean squared error (AR modelle)
- MAE: mean absolute error (AR modelle)
- MPE: mean percentage error
- MAPE: mean absolute percentage error (AR modelle)
- ACF1: first-order autocorrelation coefficient (ACF1)

The WINNER: AR Model with Order 21.

Conlusion: This conclusion is just the first steps analytics and the DB company need further inspections.

- The best situation to buy the infrastructures is the end of the odd years. For the fall of the home sales.
- The strong correlation between home sales and train ticket sales based on the number of society. It need also further analysis.
- There is a situation that the home sales are extreme falling down. If DB Company predicts precisely the factors, it can be a huge difference on budget or profit (infrastructures budget and pricing strategy)

6.5.7 Sources

- Lectures Notes Time Series Analysis [68]
- Forecasting: principles and practice [69]

Chapter 7

References & Appendix

7.1 References

[1] Columbus, Louis. (2017, December 24). 53% Of Companies Are Adopting Big Data Analytics. Retrieved from forbes.com (https://bit.ly/2TgqgWC).

[2] Paul Blase, Dan DiFillipo. (2014, August). Gut & gigabytes. Capitalising on the art & science in decision making. PricewaterhouseCoopers LLP. Retrieved from pwc.com (https://pwc.to/1rIGN6P).

[3] Investor Relations site Deutsche Bahn. Retrived from DB site (https://bit.ly/2Od4MZP)

[4] Digital Spirit. The Customer Magazine of DB Systel . Retrived from DB Systel Digital Magazine (https://bit.ly/2FjZyrg)

[5] PwC team. (2019). PwC Transport & Logistics Trend Book 2019. Retrived from Pwc Research and Insight (https://pwc.to/2SMdBe9)

[6] Knezevic, Irena. (2016, June). Big Data in food and agriculture. Retrieved from researchgate.net (https://bit.ly/2W3zTd0)

[7] Ribarics, Pal. (2016, July). Big Data and its impact on agriculture. Retrieved from researchgate.net (https://bit.ly/2HJHoSn)

[8] Sjaak Wolfert, Lan Ge, Cor Verdouw, Marc-Jeroen Bogaardt. Big Data in Smart Farming – A review. Retrived from Wageningen University & Research (https://bit.ly/2Cmn8Tt)

[9] Rands, Kevin. (2017, November 01). 4 ways big data analytics is disrupting the agriculture industry. Retrived from cio.com site (https://bit.ly/2TiFzhg)

[10] Deloitte team United Kingdom. Opportunities for analytics in the automotive industry. Retrived from deloitte.com site (https://bit.ly/2ukiD7f)

[11] Bobriakov, Igor. (2018, July 22). Top 10 Data Science Use Cases in Retail. Retrived from medium.com site (https://bit.ly/2TmWn7a)

[12] Lebied, Mona. (2018, April 5). The Impact of Big Data on The Retail Sector: Examples And Use-Cases. Retrived from datapine.com blog (https://bit.ly/2qBVGv0)

[13] Hitchcock, Erin. (2018, February 27). Five Big Data Use Cases for Retail. Retrived from datameer.com blog (https://bit.ly/2DqWcD8)

[14] Collaborative filtering. Retrived from wikipedia.com (https://bit.ly/2Iqf6Ku)

[15] Nandi, Manojit. (2017, July 14). Recommender Systems through Collaborative Filtering. Retrived from dominodatalab.com blog (https://bit.ly/2CjzCYY)

[16] SAS Warranty Analysis. Reduce warranty costs and improve product quality and brand reputation. Retrived from sas.com (https://bit.ly/2HwuDLx)

[17] Fedak, Vladimir. (2018, May 29). Big Data analytics in the banking sector. Retrived from medium.com (https://bit.ly/2JqP4LG)

[18] Evry IT companies Whitepaper. Big data in banking for marketers. How to derive value from big data. Retrived from evry.com (https://bit.ly/22CwvmW)

[19] Amakobe, Moody. (2015, July). The Impact of Big Data Analytics on the Banking Industry. Retrived from researchgate.net (https://bit.ly/2OemVXd)

[20] PwC Whitepaper Research. (2018, January). Data Analytics in the Financial Services Industry. Retrived from pwc.com (https://pwc.to/2UIcc9p)

[21] Staci, Kristel. (2018, July 6). The Powerful Role of Big Data In The Healthcare Industry. Retrived from smartdatacollective.com (https://bit.ly/2KFZlnu)

[22] Lebied, Mona. (2018, July 18). 12 Examples of Big Data Analytics In Healthcare That Can Save People. Retrived from datapine.com (https://bit.ly/2z9PIYx)

[23] Kovacevic, Andrej. (2018, November 7). Police Are Using Big Data To Predict Future Crime Rates. Retrived from smartdatacollective.com (https://bit.ly/2OyMOPM)

[24] Babuta, Alexander. Royal United Services Institute for Defence and Security Studies (RUSI). (2017, September 7). Big Data and Policing An Assessment of Law Enforcement Requirements, Expectations and Priorities. Retrived from rusi.org (https://bit.ly/2CvvMyP)

[25] Smart Vision Europe team. What is the CRISP-DM methodology?. Retrived from sv-europe.com (https://bit.ly/2Jz5yBJ)

[26] Tutorialspoint. Big Data Analytics Tutorial. Retrived from tutorialspoint.com (https://bit.ly/2HAph25)

[27] Watt, Adrienne. Open Textbook, Pressbook. Database Design. Retrived from opentextbc.ca (https://bit.ly/2TTnAmU)

[28] Microsoft Docs. (2018, May 2). Multidimensional models. Retrived from docs.microsoft.com (https://bit.ly/2HQL0Cm)

[29] Microsoft Docs. (2018, December 20). Microsoft Intune Data Warehouse data model. Retrived from docs.microsoft.com (https://bit.ly/2Fs0rxF)

[30] University of Utah. Understanding Star Schemas. Retrived from gkmc.utah.edu (https://bit.ly/2WibCA0)

[31] Gurobi Optimization Resources. Descriptive, Predictive and Prescriptive Analytics. Retrived from gurobi.com (https://bit.ly/2FvL8Wf)

[32] Verborgh, Ruben. (2013, September 10). Using OpenRefine. Packt Publishing.

[33] Atima , Han Zhuang, Ishita Vedvyas, Rishikesh Dole. (2013, December). Tutorial: OpenRefine. Retrived from University of Maryland https://casci.umd.edu/ (https://bit.ly/2JARLe1)

[34] Edureka blog. Retrived from edureka.com (https://bit.ly/2txmHjQ)

[35] Garrett Grolemund, Hadley Wickham. (2017, January 31). R for Data Science. Retrived from r4ds.had.co.nz (https://bit.ly/2HDvTN4)

[36] Edwin de Jonge, Mark van der Loo. (2013). An introduction to data cleaning with R. Retrived from cran.r-project.org (https://bit.ly/1Q4UXIQ)

[37] Kassambara, Alboukadel. Identify and Remove Duplicate Data in R. Retrived from datanovia.com (https://bit.ly/2Cxj8iY)

[38] Wishart, Jessica. (2019, January 11). 5 Reasons Why You Need The Right KPIs. Retrived from rhythmsystems.com (https://bit.ly/2HOaErg)

[39] Karlson, Karola. (2016, November 29). What Is a KPI? (Complete Guide). Retrived from scoro.com (https://bit.ly/2JFdnG9)

[40] Corporate Finance Institute Inc. Operating Cash Flow. Retrived from corporatefinanceinstitute.com (https://bit.ly/2OmVvyl)

[41] Corporate Finance Institute Inc. Current Ratio Formula. Retrived from corporatefinanceinstitute.com (https://bit.ly/2utonM4)

[42] InvestingAnswers Inc. Quick Ratio. Retrived from investinganswers.com (https://bit.ly/2JE4b4I)

[43] Kenton, Will. (2018, Maret 21). Burn Rate. Retrived from investopedia.com (https://bit.ly/2McMDZE)

[44] Corporate Finance Institute Inc. Net Profit Margin. Retrived from corporatefinanceinstitute.com (https://bit.ly/2Fzup49)

[45] Corporate Finance Institute Inc. Working Capital. Retrived from corporatefinanceinstitute.com (https://bit.ly/2TeSN2T)

[46] Corporate Finance Institute Inc. Current Accounts Receivable. Retrived from corporatefinanceinstitute.com (https://bit.ly/2TeSN2T)

[47] Corporate Finance Institute Inc. Current Accounts Payable. Retrived from corporatefinanceinstitute.com (https://bit.ly/2TeSN2T)

[48] MyAccountingCourse. Inventory Turnover Ratio. Retrived from myaccountingcourse.com (https://bit.ly/2ATQZT3)

[49] Accounting Tools. (2017, November 15). Budget Variance. Retrived from accountingtools.com (https://bit.ly/2HXGS38)

[50] Cambridge Dictionary. Sales Growth. Retrived from dictionary.cambridge.org (https://bit.ly/2Ft21zs)

[51] Grant, Mitchell. (2019, February 28). Days Sales Outstanding – DSO. Retrived from investopedia.com (https://bit.ly/2PUmJQY)

[52] Hans-Ulrich Krause, Dayanand Arora. (2010). ControllingKennzahlen Key Performance Indicators Zweisprachiges Handbuch Deutsch/Englisch. Oldenbourg Wissenschaftsverlag GmbH

[53] Steil, Tamila. (2017, June 29). The 6 Customer Service KPIs You Should Be Tracking. Retrived from userlike.com (https://bit.ly/2CbZGZk)

[54] Geckoboard. First Response Time. Retrived from geckoboard.com (https://bit.ly/2Fz7Gp0)

[55] W. Weber, Winfried. (2016, August 12). Germany's Midsize Manufacturers Outperform Its Industrial Giants. Retrived from hbr.com (https://bit.ly/2bei5oQ)

[56] Stillwagon, Amanda. (2015, March 5). 14 Key Performance Indicators (KPIs) to Measure Customer Service. Retrived from smallbiztrends.com (https://bit.ly/2YoHbK9)

[57] Designing Buildings Wiki. (2018, November 21). Cost performance index (CPI). Retrived from designingbuildings.co.uk (https://bit.ly/2TIRLIk)

[58] Roseke, Bernie. (2017, October, 20). Cost Performance Index – Earned Value Management. Retrived from projectengineer.net (https://bit.ly/2uCBbje)

[59] Usmani, Fahad. (2019, January 23). Schedule Performance Index (SPI) & Cost Performance Index (CPI). Retrived from pmstudycircle.com (https://bit.ly/2cluCMA)

[60] Factorial HR. (2017, May 19). 7 KEY INDICATORS OF HUMAN RESOURCES – HR KPI. Retrived from factorialhr.com (https://bit.ly/2mmL3th)

[61] Sisense. INTERACTIVE DASHBOARD EXAMPLES. Retrived from sisense.com (https://bit.ly/2I2Razo)

[62] KPI Dashboards. KPI Examples - Performance management starts with figuring out what to measure. Retrived from kpidashboards.com (https://bit.ly/2CK3vok)

[63] Kaggle dataset. Human Resource Analytics Dataset. Retrived from kaggle.com (https://bit.ly/2HQZ92U)

[64] Singh Dhall, Guryash. (2018, February 23). HUMAN RESOURCE ANALYSIS. Retrived from rpubs.com (https://bit.ly/2TKJsMj)

[65] Ram J, Ragul. HR Analytics- exploration and modelling with R. Retrived from kaggle.com (https://bit.ly/2WzPvoP)

[66] Mayhew, Ruth. (2019, March 01). What Is the Meaning of Attrition Used in HR?. Retrived from smallbusiness.chron.com (https://bit.ly/2V7tVYD)

[67] Reddy, Mahidhar. (2015, November 24). Top 5 Reasons for Employee Attrition & How to deal with it. Retrived from linkedin.com (https://bit.ly/2HOK0PU)

[68] Penner, Irina. (Sommer Course 2018). Lectures Note Time Series Analysis. HTW Berlin

[69] Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. Retrived from OTexts.com/fpp2.

[70] Purchasecontrol. Spend Analysis 101: How, Why, and What To Do with the Data. Retrived from purchasecontrol.com (https://bit.ly/2TTg5aN)

[71] Sievo. Spend Analysis 101. Retrived from sievo.com (https://bit.ly/2TV16wR)

[72] Microsoft Docs. (2018, Juny 06). IT Spend Analysis sample for Power BI. Retrived from docs.microsoft.com (https://bit.ly/2KchBVQ)

[73] Microsoft Docs. (2018, August 23). Add visualizations to a Power BI report. Retrived from docs.microsoft.com (https://bit.ly/2FZwwyA)

[74] Microsoft Docs. (2019, March 29). DAX function reference. Retrived from docs.microsoft.com (https://bit.ly/2GjxzKl)

[75] Microsoft Docs. (2018, Juny 06). Human Resources sample for Power BI. Retrived from docs.microsoft.com (https://bit.ly/2Kc5Rmp)

[76] Microsoft Docs. (2018, Juny 06). Opportunity Analysis sample for Power BI. Retrived from docs.microsoft.com (https://bit.ly/2UCqaxd)

7.2 Appendix

List of Figures:

[1] Columbus, Louis. (2017, December 24). 53% Of Companies Are Adopting Big Data Analytics. Retrieved from forbes.com (https://bit.ly/2TgqgWC):

Retrived from sisense.com (https://bit.ly/2Bx9NWr):

- Figure 2.8: How Big is Big Data

Retrived from Internet Live Statistics (http://www.internetlivestats.com/):

- Figure 2.9: Example of Various Online Data (Number of Websites, Social Media, Video, etc)
- Figure 2.10: Example of Various Online Data (Computers, Smartphones, Internet Traffics, ect)

Retrived from World Parameter (http://www.worldometers.info/):

- Figure 2.11: World Population Data
- Figure 2.12: Health Data

Retrived from IBM (https://ibm.co/2xf9YEi):

- Figure 2.13: The Four V's of Big Data

Retrived from Deutsche Bahn site (https://bit.ly/2Od4MZP):

- Figure 2.14: Basic Understanding of DB Group
- Figure 2.15: DB Group organization chart
- Figure 2.16: DB Group business model

Retrived from Amazon AWS site (https://amzn.to/2Co2Ada):

- Figure 2.17: Data Lake
- Figure 2.18: Data Lakes compared to Data Warehouses

Retrived from DB Systel Magazine site (https://bit.ly/2FjZyrg):

- Figure 2.19: DB Big Data Lake

[5] PwC team. (2019). PwC Transport & Logistics Trend Book 2019. Retrived from Pwc Research and Insight (https://pwc.to/2SMdBe9):

- Figure 2.20: Lack of digital culture and training
- Figure 2.21: The Technology, The impact, and The uncertainties
- Figure 2.22: The five forces transforming transport and logistics and their key driving trends

Retrived from whatsthebigdata.com (https://bit.ly/2zHjLSP):

- Figure 2.23: Data Science in Agriculture (Infographic)

[10] Deloitte team United Kingdom. Opportunities for analytics in the automotive industry. Retrived from deloitte.com site (https://bit.ly/2ukiD7f):

- Figure 2.24: Strategy, Governance, Architecture & Data Quality

Retrived from sqream.com (https://bit.ly/2U0MgJe):

- Figure 2.25: Big Data in Retail Industry

Retrived from medium.com (https://bit.ly/2TmWn7a):

- Figure 2.26: Recommendation engines

Retrived from wikipedia.com (https://bit.ly/2Iqf6Ku):

- Figure 2.27: Collaborative Filtering:

Retrived from medium.com (https://bit.ly/2Anc0E4):

- Figure 2.28: Content-based recommender system

Retrived from medium.com (https://bit.ly/2TmWn7a):

- Figure 2.29: Fraud detection

Retrived from medium.com (https://bit.ly/2TmWn7a):

- Figure 2.30: Customer sentiment analysis

Retrived from ibmbigdatahub.com (https://ibm.co/2UNMuR8):

- Figure 2.31: Big Data in Healthcare

Retrived from hitconsultant.net (https://bit.ly/2Froiiy):

- Figure 2.32: Infographic: Big Data Analytics in Healthcare

Retrived from wikipedia.com (https://bit.ly/2Froiiy):

- Figure 3.1: Cross-industry standard process for data mining

Retrived from wikipedia.com (https://bit.ly/2uqR4cx):

- Figure 3.2: Star schema

Retrived from wikipedia.com (https://bit.ly/2U3tW2p):

- Figure 3.3: Snowflake schema

Retrived from gurobi.com (https://bit.ly/2FvL8Wf):

- Figure 3.4: Big Data Analytics

Retrived from info.microsoft.com (https://bit.ly/2DFwLfD):

- Figure 3.5: Gartner Magic Quadrant Business Intelligence Tools

Retrived from powerpivotpro.com (https://bit.ly/2Fm05ZA):

- Figure 3.6: Power Query

Retrived from powerpivotpro.com (https://bit.ly/2Wm7XBo):

- Figure 3.7: Power Pivot

Retrived from confluence.atlassian.com (https://bit.ly/2UTgD1w):

- Figure 3.8: Github Version Control 1

Retrived from helloworldseries.com (https://bit.ly/2Tsb8W3):

- Figure 3.9: Github Version Control 2

Retrived from r4ds.had.co.nz (https://bit.ly/2HDvTN4):

- Figure 4.1: Tidy Data

Retrived from `md.pattern(BostonHousing)` function:

- Figure 4.2: Pattern or missing values in data

Retrived from datanovia.com (https://bit.ly/2Cxj8iY):

- Figure 4.3: Remove Duplicate Data in R

Retrived from `plot_outlier(nycflights13::flights)` function:

- Figure 4.4: Outlier Diagnosis Plot

Retrived from `plot_normality(ISLR::Carseats,Sales, CompPrice)` function:

- Figure 4.5: Normality Diagnosis Plot

Retrived from `plot_correlate(ISLR::Carseats)` function:

- Figure 4.6: Correlation Matrix

Retrived from `plot_str(ISLR::Carseats)` function:

- Figure 4.7: Data Structure

Retrived from `plot_missing(ISLR::Carseats)` function:

- Figure 4.8: Missing Rows

Retrived from `plot_histogram(ISLR::Carseats)` function:

- Figure 4.9: Frequency

Retrived from `plot_density(ISLR::Carseats)` function:

- Figure 4.10: Density

Retrived from `plot_correlation(ISLR::Carseats)` function:

- Figure 4.11: Correlation Meter

Retrived from `plot_bar(ISLR::Carseats)` function:

- Figure 4.12: Frequency Bar

Retrived from angloamericanoffice.com (https://bit.ly/2ur105T):

- Figure 5.1: Balance Sheet

Retrived from corporatefinanceinstitute.com (https://bit.ly/2OmVvyl):

- Figure 5.2: Amazon's 2017 annual report

Retrived from userlike.com (https://bit.ly/2CbZGZk):

- Figure 5.3: Customer Satisfaction Score (CSAT)

Retrived from userlike.com (https://bit.ly/2CbZGZk):

- Figure 5.4: Net Promoter Score (NPS)

Retrived from geckoboard.com (https://bit.ly/2Fz7Gp0):

- Figure 5.5: First Response Time

Retrived from userlike.com (https://bit.ly/2CbZGZk):

- Figure 5.6: Employee Engagement

Retrived from sisense.com (https://bit.ly/2U8p8JO):

- Figure 5.7: Call Center Representative Efficiency

Retrived from sisense.com (https://bit.ly/2HMWYO7)

- Figure 5.8: Profit Margin Analysis

Retrived from sisense.com (https://bit.ly/2OAXy1V)

- Figure 5.9: Ad Platform Optimization

Retrived from Microsoft Power BI Software wit Use Case Dataset Spend Analysis-IT dept.

- Figure 6.1 - 6.25

Retrived from Microsoft Power BI Software wit Use Case Dataset Human Resources sample.

- Figure 6.26 - 6.42

Retrived from Microsoft Power BI Software wit Use Case Dataset Sales Opportunity.

- Figure 6.43 - 6.56

Retrived from anomaly.io (https://bit.ly/2FHjWmj)

- Figure 6.57 and Figure 6.58: Time Series Component

Self-made graphic inspired by analyticsvidhya.com (https://bit.ly/2c6wW4Q)

- Figure 6.59: Predictions Workflows

Retrived from Penner, Irina. (Sommer Course 2018). Lectures Note Time Series Analysis. HTW Berlin

- Figure 6.60: White Noise Test

Retrived from `sarima(sales_ts_ohne_na,10,0,0)$AIC` function, `astsa` package

- Figure 6.61: AIC Model (10,0,0)

Retrived from `sarima(sales_ts_ohne_na,15,0,1)$AIC` function, `astsa` package

- Figure 6.62: AIC Model (15,0,1)

Retrived from `sarima(sales_ts_ohne_na,15,0,1)$AIC` function, `astsa` package

- Figure 6.63: AIC Model (21,0,1)

Retrived from `sarima(sales_ts_ohne_na,10,0,0)$BIC` function, `astsa` package

- Figure 6.64: BIC Model (10,0,0)

Retrived from `sarima(sales_ts_ohne_na,15,0,1)$BIC` function, `astsa` package

- Figure 6.65: BIC Model (15,0,1)

Retrived from `sarima(sales_ts_ohne_na,15,0,1)$BIC` function, `astsa` package

- Figure 6.66: BIC Model (21,0,1)

Retrived from `plot(dec.pass)` function, `forecast` package

- Figue 6.67: Decomposition of time series

Retrived from `forecast` package

- Figue 6.68: Plot decomposition of time series

Retrived from `plot(forecast(dec.pass), main="Forecast stl")` function, `forecast` package

- Figue 6.69: Plot Forecast

Retrived from `forecast` package

- Figue 6.70: Plot Forecast with Confident Interval

Retrived from `sarima.for(sales_ts_ohne_na,5,1,0,0)` function, `astsa` package

- Figure 6.71: Plot Forecast from astsa package

Chapter 8

Statement of originality / Eidesstattliche Erklärung

I hereby confirm that I have written the accompanying thesis by myself, without contributions from any sources other than those cited. This applies also to all graphics and images included in this thesis.

Berlin, 11.04.2019

Ich erkläre hiermit an Eides statt, dass

- ich die vorliegende wissenschaftliche Arbeit selbständig und ohne unerlaubte Hilfe angefertigt habe,
- ich andere als die angegebenen Quellen und Hilfsmittel nicht benutzt habe,
- ich die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe,
- die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfbehörde vorgelegen hat.

Berlin, 11.04.2019