1 Introduction

1.1 Name the three main types of analytics tasks in business intelligence, as well as the question each of them addresses.

1.2 To which analytics task does the following belong to:

- optimization:
- data mining:
- data warehousing:
- forecasting:
- simulation:

1.3 Data science is the defined as intersection of which three competences?

1.4 What is the difference between OLTP (online transaction processing) and OLAP (online analytical processing)? In particular, consider their use, the way they organize their data, and the objective they are optimized for.

1. use:
2. data organization:
3. optimization objective:

1.5 Name the three main Vs in big data and explain them.

2 Descriptive Analytics Part 1

2.1 Which of the following data content types is considered structured?

( ) text in product reviews

( ) ratings from one to five stars in product reviews

( ) counts of the word "good" in a review

( ) product category, encoded in numbers

( ) product category, encoded as "books","films","shoes",...

( ) product images

( ) transaction value

2.2 Indicate for which level of measurement (1) mode, (2) median, (3) arithmetic mean, (4) geometric mean are appropriate measures

- nominal:
- ordinal:
- interval:
- ratio:

2.3 In data streams, data is

( ) available at once

( ) arrives sequentially

2.4 What are the characteristics of streaming data? Can you relate them to the big data "Vs"? What are the corresponding challenges?

2.5 Consider the example of calculating the average sales per month, where data from the last five years is available. Using this example, explain why it is important to consider the temporal ordering of data instances.

2.6 Name and describe four data quality and usability metrics

2.7 Review Application Case 2.2 and the questions therein.

2.8 What is class imbalance? What is the problem of measuring a classifier's accuracy under extreme class imbalance? Which counter-measures could be used in preprocessing?

2.9 Describe the approach provided in Application Case 2.2 for assessing the impact (importance) of a variable in the classification model.

2.10 Name the four steps of data preprocessing. Name one example for each.

References

[Sharda et al., 2018] Sharda, R., Delen, D., and Turban, E. (2018). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. Pearson, 4 edition.

3 Descriptive Analytics Part 2

3.1 Probability Density (PDF) and Cumulative Distribution (CDF) Functions

You are working for a Dutch-based, international group of retailers named SPQR, who operates stores across 40 countries world-wide. From a previous analysis of historical data, the company knows that the monthly aggregated sales (in Euros) of its product X are distributed according to a Gaussian normal distribution with mean $\mu = 230,000$ and standard deviation $\sigma = 20,000$. For the following questions, assume stationarity, i.e. that the distribution does not change.

1. SPQR asks you to provide an estimate of the probability that the sales will exceed $200,000$ next month.
   Hint: Use the function $pnorm(q, mean, sd)$ in R.

2. SPQR asks you to provide an estimate of the probability that the sales will be between $200,000$ and $260,000$ next month.
   Hint: Use the function $pnorm(q, mean, sd)$ in R.

3. Using the probability density function $dnorm(x, mean, sd)$, provide estimates of the density (i.e., for the probability that the next month's sales are in an infinitesimal interval around these X values) for (a) $X = 230,000$ and (b) $X = 240,000$. Which one is greater?

3.2 Using R for calculating selected statistics:

For the following three samples

- $X_1 = \{-2, -1, -.5, 0, 0.25, 0.25, 1, 2\}$
- $X_2 = \{0, 0.1, 0.3, 0.7, 1.5, 3, 7\}$
- $X_3 = \{-2, -1.25, -1, -.8, -.5, 0, 0, 0\}$

Use R to calculate for the following statistics

- Mode $modeval(X)$ (see hint below)
- Median: $median(X)$
- 1st quantile: $quantile(x = X, probs = .25)$
- 3rd quantile: $quantile(x = X, probs = .75)$
- Arithmetic Mean: $mean(X)$
- Minimum, Maximum: $min(X), max(X)$
- Range: $max(X) - min(X)$
- Interquartile Range IQR: $quantile(x = X, probs = .75) - quantile(x = X, probs = .25)$
- (Unbiased) Variance: $var(X)$
- (Unbiased) Standard Deviation: $sd(X)$ or $var(X)^{.5}$
- (Sample) Skewness: $skewness(X)$

Hint: Load and use the library moments and use the following function definition for the mode:

```
library(moments);
# defining a function modeval(X) to calculate the mode of X:
modeval = function(X){
  return(as.numeric(names(sort(-table(X))[1])));
}
```

3.3 Kullback-Leibler Divergence

The SPQR group is interested in analysing the relative importance (i.e., share of total sales) of its different product categories (for simplicity, we consider three main categories A,B,C) to its total sales (world-wide, and in selected countries). These shares are given in Table 1. You are tasked to compare, how similar the distribution of sales' shares is in the Netherlands Q_1 (resp., Italy Q_2) to the aggregated (world-wide) share distribution (P), by calculating the Kullback-Leibler Divergence between P and Q_1 (resp., Q_2).

	A	**B**	**C**
World-Wide, P	0.4	0.3	0.3
Netherlands Q_1	0.4	0.35	0.25
Italy Q_2	0.5	0.3	0.2

Table 1: Shares of product categories in sales.

Note: You can use the following function definition for the KL-Divergence in R:

```
# definition of a function kldiv(p,q) to calculate the KL-Div(p||q)
kldiv = function(p,q){
  frac = p/q;
  plogfrac = p*log(frac);
  plogfrac[p == 0] = 0;
  plogfrac[(p != 0)&(q == 0)] = Inf;
  return(as.numeric(sum(plogfrac)));
}
```

3.3.1 Independence

For the SPQR group in the example above, the sales' shares by product category and country are given in Table 2. If one interprets country and product as random variables X and Y, how would this table look like if country X and product Y were independent of each other?

Product Y:	**A**	**B**	**C**	**Total**
Country X:				
Netherlands Q_1	0.24	0.21	0.15	0.6
Italy Q_2	0.2	0.12	0.08	0.4
Total	0.44	0.33	0.23	1.0

Table 2: Sales' shares by product category and country.

3.3.2 Relationship between Variables

The SPQR group wants you to analyse a possible relationship between the sales of two of its products, X, and Y, with their sales data given in Table 3 below: Use R to calculate the covariance $cov(X, Y)$, Pearson's linear

Product	**1**	**2**	**3**	**4**	**5**
X	0.1	0.4	0.5	0.8	0.9
Y	0.9	0.3	0.2	0.21	0.1

Table 3: Sales data for products X and Y

correlation coefficient $cor(X, Y, method = \text{“}pearson''$), and Spearman's rank correlation coefficient $cor(X, Y, method = \text{“}spearman''$). Furthermore, use lm to fit a linear model with X being the explanatory variable, and Y being the dependent (or response) variable.

3.4 Streaming or Time Series Data

The SPQR group's data warehouse provides (aggregated) sales records (in millions of Euros) of several quarters (see Table 4). You are called to settle the dispute between two colleagues, one who argues that all data should be used to calculate the average sales of this product, the other arguing to use just the most recent data. Use the arithmetic mean to demonstrate the effect of (a) using all data vs. (b) using a windowing approach that divides the data into three chunks, one for each year.

Year				**2015**				**2016**				**2017**
Quarter	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Quantity	0.3	0.4	0.3	0.35	0.55	0.6	0.5	0.6	0.4	0.5	0.45	0.4

Table 4: Sales data (in millions of Euros) for products X over time

3.5 Indicate which of the following statements are correct

( ) Data (or Information) Visualisation and Visual Analytics are precisely the same.

( ) Information Visualisation is foremost retrospective and descriptive, answering the questions such as what has happened or is happening.

( ) Information Visualisation is foremost predictive, future-oriented, answering the questions such as what will happen.

( ) Predictive Analytics is foremost predictive, future-oriented, answering the questions such as what will happen.

( ) Visual Analytics is the combination of information visualisation and predictive analytics.

3.6 Indicate which of the following statements about narrative visualization are correct

( ) The Martini Glass schema corresponds to (a) starting with a broad view that allows the user to interactively select among various aspects potential points of interest, and (b) then successively narrowing the scope down (e.g., by focusing on a single KPI).

( ) Quite the opposite! The Martini Glass schema starts with a narrow focus point and allows initially little interaction. Then, the scope is subsequently widened and more and more interaction is allowed.

( ) The Drill-Down-Story starts with a "map" that outlines potential points of interest, and allows the user to interactively select and explore aspects in depth.

( ) The Drill-Down-Story starts with one focus point, and leads the user through a step-by-step analysis, thereby "drilling down" to the underlying facts.

( ) The Interactive Slideshow is a an approach that structures the information into several subsequent steps ("tabs"), and allows the user to navigate through them interactively at their own speed.