

**KAN-CBUSV2036U**  
**Applying Data Analytics in Digital Business (B)**  
**EXAM – Spring 2021**

This is the exam for *Applying Data Analytics in Digital Business (B)*. On the following pages you find instructions for the report that you are asked to compile as well as some more general tips to take into account when writing an academic report.

Your hand-in can consist of a maximum of 10 pages in total, please check the rules at [myCBS.dk](http://myCBS.dk). There will be no further explanation or supervision during the examination period.

This 10-page report includes an answer sheet including your answers to all questions and one DO file including all Stata codes you have used in answering questions.

Please make sure to take the following general suggestions into account when compiling your report:

- Cite all external material that you use in the process of compiling your exam report and provide a bibliography in alphabetical order
- Do not copy and paste any text or other content such as figures from external sources without properly citing it
- Make sure to hand in an original report that was produced by you individually
- Structure your report in a way that allows the reader to easily follow your arguments
- Make sure to check your final report for completeness before you hand it in
- Make sure to hand the exam in before the deadline
- 

Good luck with the exam!

## **Section 1: Survey Research**

### **Question 1:**

Amazon, as the e-commerce giant, has seen massive volumes of sales every quarter, with values exceeding \$100 billion each time (Amazon.com, 2022) since the pandemic. Third-party sellers tend to benefit tremendously from Amazon's marketplace affiliation.

You are working in the marketing department for a company which is a luxury audio-visual system manufacturer. The company has its own offline and online advertising and retailing channels. Now the managers are thinking about whether to go to Amazon or not. Your department is responsible for the market investigation.

There are clearly pros and cons to joining Amazon. The platform can expand your customer base, at the same time, it's a constant price war on the platform where a cheaper alternative will always exist in a landscape of thousands of different sellers.

Your department plans to investigate the influence of opening a channel on Amazon on the existing customers and potential new customers. Please discuss how you will design a survey that can help collect data to support the decision making based on the following requirements.

1. Your survey should include at least two related constructs, e.g., brand loyalty, brand image, etc, in your survey. Your measures can be originated from existing scales in literature but must be adapted to this research context. Please include the definitions and measures as well as the references in the report.
2. Your target population should include both existing customers and potential new customers. Please define your target population and sampling frames and choose the sampling methods. Also please report your plan of survey administration.
3. Design your survey on Qualtrics and provide the link in the report. Your survey should include the introduction, some demographic questions, possible filtering questions and all the items for the constructs. Please make sure that the link is open for access.

### **Question 2:**

You are running a startup with your business partners. You have launched a mobile app of sport fan community and the name of the App is Sporty. On the app, news can be posted, and the sport fans can like, repost, and comment on the posts as well as communicate with each other. Some additional services can be purchased by the customers.

The 1<sup>st</sup> version of the App has been in the market for half a year, you would like to investigate what factors will influence users' intention and behavior to use the App and then see what you can improve in the App. Therefore, you recruited some existing customers and asked them to complete a survey.

You designed the survey according to the theory of UTAUT2<sup>1</sup> and adapted the research model to your business context. The research model and the construct scales are included in this document.

Some data have been collected. Please analyze the data file **question2.dta** and answer the following questions. Please write up the results and interpret the results as you are writing the data analytics part of a paper. Include all necessary tables and figures. If there is not enough space, please put the figures and tables in the appendix. (*Submit all Stata codes used to answer the following questions in the DO file*)

**Question 2.1.** In general, do customers use the app Sporty frequently? Get your conclusions based on the descriptive statistics from the three items of *Use Behavior*. Plot some graphs if needed.

**Question 2.2.** Are the measures of the constructs *Subjective Norm* and *Facilitating Conditions* reliable? How about the validity?

**Question 2.3.** According to the theoretical model, you are going to test the following hypotheses. Please report your analysis result and conclusion.

*H1. Hedonic Motivation will have a positive effect on Behavioral Intention*

*H2: Price Value will have a positive effect on Behavioral Intention*

*H3. Subjective Norm will have a positive effect on Behavioral Intention*

*H4. Behavioral Intention will have a positive effect on Use Behavior.*

*H5. Facilitating Conditions will have a positive effect on Use Behavior.*

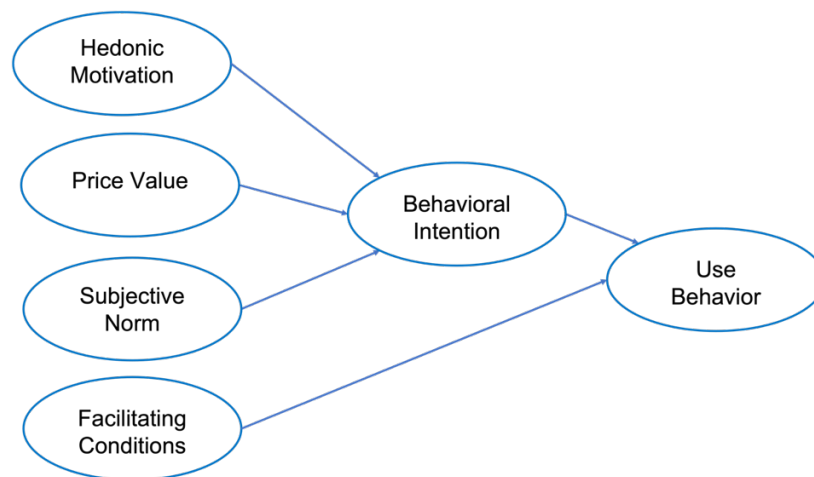


Figure 1. Research Model

Measurement scales for the mobile App Sporty

Hedonic Motivation (7-point Likert scale anchored with strongly disagree to strongly agree)

HM1. Using the mobile App Sporty is fun.

HM2. Using mobile App Sporty is enjoyable.

---

<sup>1</sup> Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, 157-178.

HM3. Using mobile App Sporty is very entertaining.

HM4. Using mobile App Sporty is very playful.

Price Value (7-point Likert scale anchored with strongly disagree to strongly agree)

PV1. Sporty is reasonably priced.

PV2. Sporty is a good value for the money.

PV3. At the current price, Sporty provides a good value.

PV4. Given your experiences of Sporty and the current price how would you rate the App?

Subjective Norm (SN) (7-point Likert scale anchored with strongly disagree to strongly agree)

SN1: Most people who are important to me think I should use Sporty

SN2: Most people who are important to me would want me to use Sporty

SN3: People whose opinions I value would prefer me to use Sporty

Facilitating Conditions (FC) (7-point Likert scale anchored with strongly disagree to strongly agree)

FC1: I have the resources and the knowledge and the ability to make use of Sporty

FC2: A central support was available to help with problem in using Sporty

FC3: The customer support provided most of the necessary help and resources for Sporty

Behavioral Intention (BI)

BI1: I predict I will continue to use Sporty on a regular basis (7-point Likert scale anchored with strongly disagree to strongly agree)

BI2: What are the chances in 100 that you will continue as a Sporty user? (1) Zero; (2) 1–10%; (3) 11–30%; (4) 31–50%; (5) 51–70%; (6) 71–90%; or (7) more than 90%

BI3: I would use Sporty rather than any other Apps available (7-point Likert scale anchored with strongly disagree to strongly agree)

Use Behavior (USE)

USE1: On an average working day, how much time do you spend using Sporty?

(1) Almost never; (2) less than 30 min; (3) from 30 min to 1 h; (4) from 1 to 2 h; (5) from 2 to 3 h; and (6) more than 3 h

USE2: On average, how frequently do you use Sporty?

(1) Less than once a month; (2) once a month; (3) a few times a month; (4) a few times a week; (5) about once a day; and (6) several times a day

USE3: How many different functions have you tried out in Sporty?

(1) None; (2) one; (3) two; (4) three; (5) four to five; and (6) six or more than six

## Section 2: Experimental Research

### Question 3:

Assume you are the CEO of a startup that wants to use digital IoT sensors to fight pollution. Your startup sells IoT sensors that send data about the CO<sub>2</sub> level in the air to national agencies. Additionally, your startup offers advisory services. You have been hired by a national agency in your country to help it understand possible ways to measure and combat air pollution. The agency has bought your sensors to measure the impact of setting different driving speed limits on the air CO<sub>2</sub> levels. The rationale is that limiting the speed limit for cars will reduce the measured CO<sub>2</sub> in the air. The agency asks you how you would design an experiment to learn if reducing the driving speed limit reduced the CO<sub>2</sub> in the air? Make sure that you follow the *Guidelines for Designing an Experiment* in your description. With which challenges might you be confronted when running such an experiment?

### Question 4:

Following your advice, the agency has asked you to set up an experiment to measure the impact of reducing the speed limit in the capital city by 10 km per hour. Your sensors have been implemented in 150 stations, distributed randomly all over the country. Your sensors measure the level of CO<sub>2</sub> in a point that is close to a road. Note that there are different speed limits across the streets. Hence, the measuring stations are located in streets whose speed may vary. You measure the level of CO<sub>2</sub> (in parts per million) before and after your client (the national agency) has limited the speed limit of the road in which the station is located by 10 km per hour. Some of the stations are located on a road in which the speed limit is above 60 km per hour. Some of the stations are located on a road in which the speed limit is below 60 km per hour.

Your client wants to learn the effect of limiting the speed limit by 10km per hour (regardless of the original speed limit of the road). Additionally, your client wants to learn if the potential impact of the speed limit reduction is different if the original speed limit of the road (i.e., the speed limit prior to the reduction) is above or below 60 km per hour. You collected the variables described in the table:

Variable	Interpretation
<i>Speedreduction</i>	<i>Speedreduction</i> = 1 indicates that the speed reduction has been implemented; <i>Speedreduction</i> = 0 indicates that the speed reduction has NOT been implemented
<i>Below60</i>	<i>Below60</i> = 1 for streets in which the original speed limit of the road is below 60 km per hour; <i>Below60</i> = 0 for streets in which the original speed limit of the road is above 60 km per hour;
<i>CO2ppm</i>	The level of CO <sub>2</sub> measured by the station in parts per million.

Load the dataset: *examcbs2022.dta* (This is a simulated data set)

Answer the following questions by carefully analyzing the given dataset and discuss your analysis findings.

(1) What is the impact of reducing the speed limit by 10 km per hour on the CO<sub>2</sub> level? (2) Is the CO<sub>2</sub> level in streets whose original speed limit is above 60km per hour different from the CO<sub>2</sub> level in streets whose

original speed limit is below 60km per hour and why? (3) Is there any difference in the impact of reducing the speed limit on streets whose original speed limit is above or below 60 km per hour and why? (4) Which difference in the CO2 level should you expect for the streets where you reduce the speed limit and whose original speed limit is above 60 km per hour, compared to the streets where you reduce the speed limit and whose original speed limit is below 60 km per hour? (5) Write a short (200 words max) summary explaining the key takeaways of your analysis for someone that knows nothing about data analysis. (*Submit all Stata codes used to answer the above questions in the DO file*)

---

### Section 3: Prediction Models

#### Question 5:

MyMarket is an online grocery store, whose key competitor is Coop.dk because of the same business model. You are recently hired by MyMarket and assigned some analytics tasks to do. You have done some work with preliminary results. Your supervisor, Adam, who does not have any technical background, needs you to interpret your analysis results for him. Please answer the following questions raised by Adam.

**Question 5-1:** You were asked to build a model to predict whether a customer will opt in the membership based on some available information. By analyzing a sample dataset, you have the following variables.

Variable name	Definition
$optin_i$	whether customer $i$ opts in MyMarket membership in 2021, 1=Yes, 0=No
$n\_of\_visits_i$	Total number of website visits by customer $i$ in 2021
$gender_i$	Gender of the customer $i$ , 1 for male; 0 for female

You built up a prediction model based on logistic regression and want to apply it to predict a customer will opt in membership service. After running the logistic regression model properly, you obtained the following estimation for two predictors.

Variable names	Estimated coefficient	Odds ratio	p-value
$n\_of\_visits_i$	0.28	1.323	0.02
$gender_i$	-1.93	0.145	0.03
Intercept	-1.255	0.285	0.001

Explain the concept of odds ratio to Adam. What is the odds ratio? Why do we need odds ratio? How can we apply the odds ratio to interpret the results from the above table? By applying your estimated coefficients, please predict whether a female customer, who visited MyMarket 22 times in 2021, will opt in membership service or not? Why?

**Question 5-2:** You also included a confusion matrix in your report for Adam. The confusion matrix was created based on the cutoff point 0.5, which is presented below.

		Model	
		$Pr(Y=1)>0.5$	$Pr(Y=0)<0.5$
Real life	Opt in ( $optin_i=1$ )	1654	123
	No Opt in ( $optin_i=0$ )	354	843

Explain to Adam about 1) the concept of confusion matrix and 2) how we can use the numbers from the above confusion matrix to evaluate the (logistic regression) model performance.

**Question 5-3:** 500 customers opted in MyMarket membership in 2021. You want to design a study to investigate their life value in 2022. The life value is defined as the time difference between 1<sup>st</sup> Jan 2022 and the opt-out date in 2022. Based on the requirement, how will you design this customer life study? Which kind of analysis or method can you apply in this study? Why?

### Question 6:

You work as a marketing analyst for the catering industry. You are assigned a task to analyze a dataset from TripAdvisor (*tripadvisor\_italian\_restaurants.csv*), which includes all Italian restaurant reviews. Each variable and its definition are given below. (Submit all Stata codes used to answer the following questions in the DO file)

Variable	Definition
restaurant_id	Identifier of restaurant
region	Region name in Italy
province	Province name in Italy
awards	List of award(s) of a particular restaurant (one restaurant may have multiple rewards, which are all recorded in one cell)
vegetarian_friendly	Whether the restaurant is a vegetarian friendly restaurant, denoted as “Y” or not, denoted as “N”
gluten_free	If the restaurant provides gluten free food, denoted as “Y”, otherwise, denoted as “N”
open_days_per_week	Number of opening days per week of the restaurant (in day)
open_hours_per_week	Number of opening hours per week of the restaurant (in hour)
overall_rating	Overall rating of the restaurant, ranging from 1 to 5 (the higher the better)
total_reviews_count	Total number of reviews to the restaurant in TripAdvisor
food	Food quality rating of the restaurant, ranging from 1 to 5 (the higher the better)
service	Service quality rating of the restaurant, ranging from 1 to 5 (the higher the better)
value	Value rating of the restaurant, ranging from 1 to 5 (the higher the better)
atmosphere	Atmosphere rating of the restaurant, ranging from 1 to 5 (the higher the better)

**Question 6.1:** Using the Stata commands to complete the following data wrangling tasks. 1) Create a new binary variable, named “award\_1”. If the variable “awards” is null, “award\_1” is equal to 0. Otherwise, “award\_1” is equal to 1. 2) Convert string categorical variables, “region” and “province” into numerical categorical variables and name the new variables as “region\_code” and “province\_code” respectively. 3) Create two new binary variables, named as “veg\_friend” and “glu\_free”. “veg\_friend” is 1 if “vegetarian\_friendly” is “Y”, and 0 otherwise; “glu\_free” is 1 if “gluten\_free” is “Y”, and 0 otherwise. 4)

Find out the Top 5 Italian regions having the most vegetarian friendly restaurants and fill the following Table.

Region name	# of vegetarian friendly restaurant

**Question 6.2:** Select predictors from the following variables, “award\_1”, “veg\_friend”, “glu\_free”, “open\_days\_per\_week”, “open\_hours\_per\_week”, “total\_reviews\_count”, “food”, “service”, “value”, and “atmosphere”, to build a model to predict overall rating of a restaurant. Report the estimated coefficients of the selected predictors and their p-values. Does your model include the variables, “atmosphere” and “award\_1”? If they are included, please provide interpretation of their impact on overall ratings. If “Not”, please justify why they are excluded.

**Question 6.3:** Your manager suggests the competition in the catering industry in the local province, measured by the total number of restaurants in the same province, may be an important predictor of overall ratings? Create a new variable, named as “competition”, measuring the total number of restaurants in the province. Include “competition” in the model you built up in Question 6.2, does your manager provide a valid suggestion or not? Why?