STAT 425 Case Study 1 (Fall 2024)

Halloween Candy Data

Instructor: A. Chronopoulou

Case Study Overview

The goal of the case study is to find what makes a Halloween candy more desirable. *FiveThirtyEight* website devised a survey to rank various Halloween candy. With a self-selected sample of `8,371` individuals, they randomly generated `269,000` match-ups between different candy. For each candy, they computed the `winpercentage` (variable `winpercent`) which represents the percent of wins of a candy in a match-up. The top 5 winners are:

Candy Name	% of Wins in a Match-up
Reese's Peanut Butter Cup	84.2%
Reese's Miniatures	81.9%
Twix	81.6%
Kit Kat	76.8%
Snickers	76.7%

To decide which are the qualities of a Halloween candy that make it desirable, we have the following variables in our disposal: `chocolate`, `fruity`, `caramel`, `peanutalmondy`, `nougat`, `crispedricewafer`, `hard`, `bar`, `pluribus`, `sugarpercent`. All the data can be found in the `candy-data.csv` file on Canvas.

Variable Name	Description
`chocolate`	Does it contain chocolate?
`fruity`	Is it fruit flavored?
`caramel`	Is there caramel in the candy?
`peanutalmondy`	Does it contain peanuts, peanut butter, or almonds?
`nougat`	Does it contain nougat?
`crispedricewafer`	Does it contain crisped rice, wafers, or a cookie component?
`hard`	Is it a hard candy?
`bar`	Is it a candy bar?
`pluribus`	Is it one of many candies in a bag or box?
`sugarpercent`	The percentile of sugar it falls under within the data set.

(*) In all cases `1` corresponds to *yes* and `0` corresponds to *no*.

In this case study, you **should**: *(i)* fit a **multiple regression model** that explains the *desirability* of a Halloween candy; *(ii)* use your model to estimate/predict the *winpercent* of at least two of your favorite candy; *(iii)* identify the qualities of the *ideal* Halloween candy.

Learning Objectives

By the end of this case study, you will

1. enhance your skills in using R for the purpose of performing a multiple linear regression.
2. independently apply the regression tools discussed in class in a real-world problem.
3. evaluate the applicability of the regression model.
4. draw conclusions, and make decisions about the initially stated research question(s).
5. interpret your statistical outcomes using plain English.
6. demonstrate your team collaboration skills.

General Case Study Guidelines

1. The case study should be done in a group of **2–4 students**. You are free to choose your own group. If you do not have a group, please use the Google form on Canvas to let me know and I will randomly assign you to a group.
   Remark: *Case studies conducted by a single student will receive no credit.* If there are extenuating circumstances that prevent you from working in a group, please contact the instructor asap.

2. Data Analysis:

   (a) You should start with an exploratory data analysis: choose appropriate summary statistics and plots to describe the data.

   (b) Start by fitting the full MLR model and reduce it to a model that only contains statistically significant variables. Non-MLR models and information-theoretic methods for model selection should **not** be considered in this case study, and if submitted will receive no credit.

   (c) Do not forget to check diagnostics. If remedial measures are needed, take appropriate actions to remedy any departures (along the lines of what we discuss in this course). If a MLR model does not seem appropriate, explain why.

   (d) The significance level α is up to you to choose, but you should select it *in the beginning*.

3. AI-generated report and/or code will receive a zero score for the case study and will be reported to FAIR.

Deliverables

The case study should be submitted on Gradescope as a group (only one case study per group) and should contain the following files:

(1) an **R Markdown**-generated HTML or PDF technical report containing all the steps in your analysis with discussion of the results along with the corresponding files. This should be professionally and clearly written addressed to someone *who knows statistics*. Do not

forget to include an introduction and a conclusion. In your report, please make sure that you remove/hide any R-generated output that is not necessary, e.g. do not print the data or the residuals.

(2) a **PDF** file containing a 3-5 **slides presentation** of your project. All members of the group will need to present **in person** and answer project-related questions. Presentation time slots will be posted on Canvas for the groups to sign up.

Grading

The grading of the case study consists of two-parts:

(1) 50% report (rubric attached)

(2) 50% presentation (rubric attached)

Deadline: Submit *one case study report and presentation per group* on Gradescope by **Monday, October 28** @ 11.59PM. Presentations will be scheduled **on or before Monday, October 28**.

I think that analyzing real data is one of the most fun challenges in Statistics!
This is just a tiny glimpse of how that looks like in practice!
Try not to overthink it and enjoy!!
Good luck!

STAT 425: Case Study 1 Report Rubric (Fall 2024)

Category	Points		
	1	2	3
Introduction	Project description is missing or it is poorly written and overly simplistic.	Project description is generic.	Project description is well motivated, interesting, and insightful.
Model Selection Process	Model selection process is missing or is incorrect.	Model selection process is relevant and correct, but there is no justification of the various steps.	Model selection process is relevant, correct, and includes many details and justification of various steps.
Diagnostics	Model assumptions are not checked. More than one missing, and there conclusions are incorrect.	Model assumptions are checked - but one is missing, and some are not correct.	All model assumptions are checked, using appropriate graphs and/or statistics. Explanation of departures is clearly discussed in detail.
Conclusion	Conclusion is missing or it is unclear/ vague.	Clear and almost complete statement of the analysis results.	Clear and complete statement of the analysis results. Conclusion is presented in both statistical terms and using layman's terms.
Report Presentation/ Organization	Presentation is not clear, the report is poorly organized and hard to read.	Presentation is good, and the report is easy to read and understand. Some graphs might not be polished.	Presentation is excellent and professional. All graphs are polished and professionally presented.
Use of R outputs/ R Code	Unrelevant R output is left in the text. There is no discussion of the R results.	R output is all necessary and relevant to the data analysis task.	R output is carefully tailored to the problem. Comments and discussion of results follows all output.

STAT 425: Case Study 1 Presentation Rubric (Fall 2024)			
Category	**Points**		
	1	**2**	**3**
Data Analysis	Group members are not able to clearly explain the steps taken to analyze the data.	Ability to clearly explain basic steps taken to analyze the data.	Ability to clearly explain the steps taken to analyze the data. Insightful comments are added and the group is able to clearly explain all steps taken efficiently and effectively.
Results/ Conclusion	Group members not able to convey conclusion or not able to explain it in layman's terms.	Ability to convey the main "take-home" messages from the analysis, supported by analysis conducted.	Ability to convey the main "take-home" messages from the analysis, supported by analysis conducted, using only necessary output and using both technical and layman's terms.
Ability to Answer Question	The group members were not able to answer the majority of the questions.	The majority of the group members are able to answer questions.	Everyone in the group was able to answer questions.
Group Interaction	Poor	Good	Excellent