Project 01

Scott Graham, Kaisa Roggeveen

January 25, 2018

1.

One of the more obvious things to look for is any cases where the patient answers "No", "Don't Remember", "Missing Answer" or "Unexpected Answer" to Question 03, and then fills out Questions 04-07, thereby ignoring the instructions.

Table 1: Patients with Errors in Questions 04-07

Patient Number	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Waiting Time
52	1	1	2	0	0	0	8	1	8	2	0
106	1	1	1	4	1	3	0	6	3	1	0
240	1	2	2	4	3	3	1	3	1	3	30

Patient 52 is most likely a case of a data entry error, as their answer to Question 08 was erroneously entered as an 8 (Missing Answer), instead of a 0 (No answer expected). Patient 240 on the other hand either filled out Question 03 incorrectly, their response was entered incorrectly, or perhaps misread what the following questions asked. This may have resulted in them filling out information about a previous visit to the clinic where they did meet Dr. Sayah. Patient 106 gave a non-zero answer to Questions 04-06, but then answered 0 to Question 07.

By looking at the types of the columns of the data, Question 01 had something entered in as a non-numeric value, which is an issue.

Table 2: Patients with Errors in Question 01

Patient Number	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Waiting Time
300	l	1	8	0	0	0	0	4	5	4	0

Patient 300 had their response to Question 01 entered in as an "l" instead of a "1".

Another issue arises for patients who indicate in Question 2 that they met a nurse immediately, but still listed a waiting time.

Table 3: Patients Who Indicated They Didn't Wait, but Listed a Waiting Time, or the Opposite

Patient Number	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Waiting Time
195	3	2	1	3	2	2	1	3	5	2	0
252	1	1	1	3	3	4	1	2	1	2	20

Patient 252 is guilty of this, and either their wait time should be set to 0, or their response to Question 02 should be changed. The opposite hold true for Patient 195, who didn't immediately meet a nurse, but had a wait time of 0.

For each question, responses can only be coded as $0, 1, \ldots k, 8, 9$, where k is the number of options given in the survey. As such, one should check to see if any patients gave responses outside of this scope.

Table 4: Patients Who Answered with a Response Outside the Scope of the Question

Patient Number	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Waiting Time
20	1	1	1	3	6	7	1	2	3	3	0

2.

Table 5: Patients with an Unexpected Answer

Patient Number	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
5	9	1	1	2	3	2	1	3	3	1
8	1	1	1	2	3	3	1	1	2	9
9	9	1	1	2	2	1	1	4	3	2
15	1	1	1	3	3	2	9	2	6	1
16	1	1	1	2	3	2	9	3	1	1
25	1	1	1	3	2	3	1	3	9	1
35	1	1	1	3	2	1	9	1	3	4
58	1	1	1	3	1	1	1	3	2	9
60	9	2	1	3	2	4	1	4	3	3
76	1	2	1	3	3	1	1	4	9	5
79	2	9	1	2	2	1	3	4	2	2
87	1	1	1	3	2	4	9	3	2	4
107	1	1	9	0	0	0	0	3	1	3
125	1	9	1	3	3	2	1	4	3	2
151	1	1	1	3	4	3	9	5	2	2
188	1	1	1	3	9	4	2	2	2	2
207	1	2	1	3	3	3	9	3	3	2
208	2	1	1	2	3	9	2	5	3	3
226	1	1	1	2	2	5	1	1	6	9
248	1	1	1	3	9	3	2	5	3	3
266	1	1	1	9	4	2	1	2	4	4
287	1	1	1	9	2	5	1	2	2	2
291	2	1	1	3	3	3	1	9	3	2

Table 6: Total Patients with an Unexpected Answer

Total Patients	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
23	3	2	1	2	2	1	6	1	2	3

3.

Table 7: Patients with a Missing Answer

Patient Number	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
7	1	1	1	2	3	5	1	2	3	8
11	8	2	1	3	1	2	2	2	1	2
52	1	1	2	0	0	0	8	1	8	2
68	1	2	1	1	3	1	3	8	1	2
77	1	1	1	2	2	3	3	3	2	8
94	2	1	1	5	3	8	1	2	3	1
97	1	8	1	3	2	2	1	2	2	2
98	1	1	1	2	8	5	1	2	3	3
120	1	1	1	3	3	8	2	1	3	3
134	8	1	1	2	3	3	2	5	4	2
165	2	1	1	2	3	8	1	2	1	4
172	2	2	8	0	0	0	0	1	5	5
227	1	3	1	2	4	2	8	8	3	3
235	1	3	1	3	6	2	1	3	8	2
243	1	1	1	3	8	1	1	1	3	3
249	1	8	1	3	3	4	1	3	2	3
267	2	1	1	2	1	3	1	2	2	8
286	1	1	1	2	3	3	8	1	2	5
299	2	1	1	3	2	4	1	3	1	8
300	1	1	8	0	0	0	0	4	5	4

Table 8: Total Patients with a Missing Answer

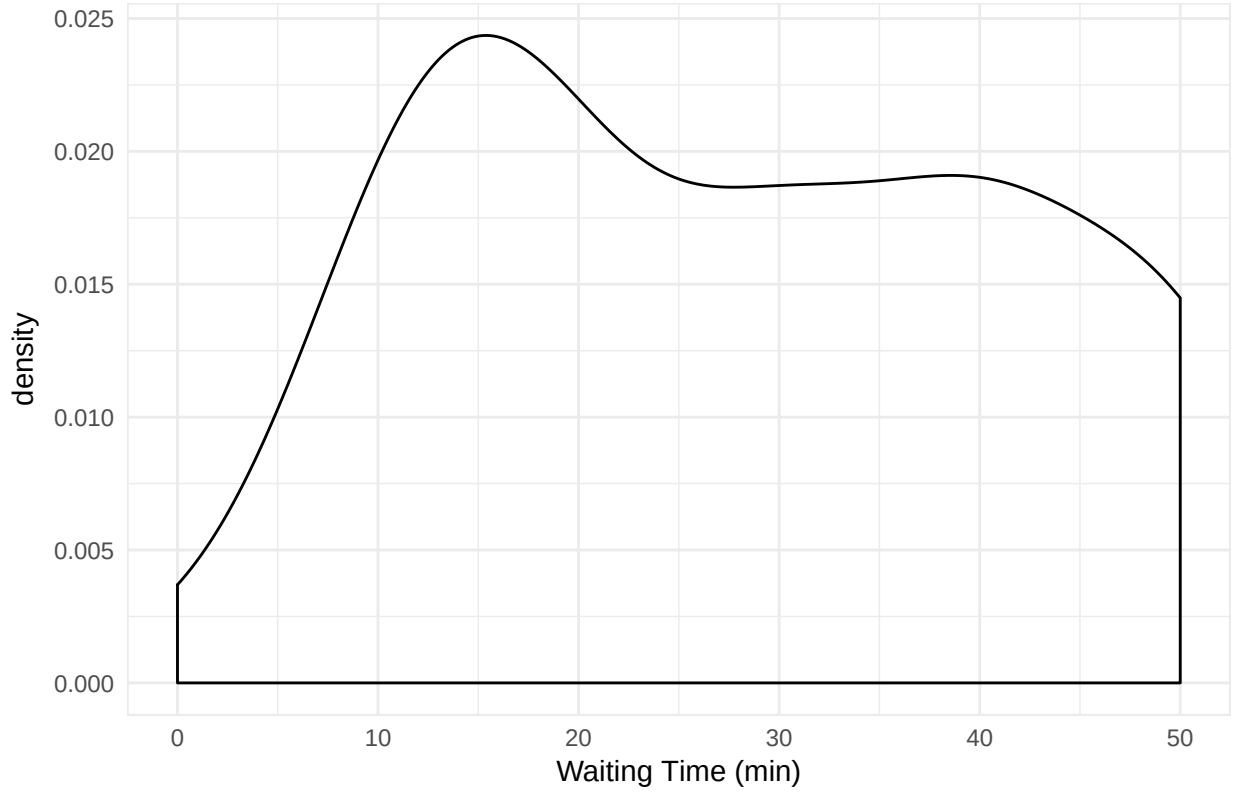
Total Patients	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
20	2	2	2	0	2	3	3	2	2	4

4.

Table 9: Summary Statistics for Waiting Time for Those Who Didn't Meet a Nurse Promptly

Waiting Time
Min. : 0.0
1st Qu.:15.0
Median :30.0
Mean :28.2
3rd Qu.:40.0
Max. :50.0

Figure 01: Kernel Density Estimator for Waiting Time



5.

For this, we considered the responses to Question 10, "Considering all aspects of your most recent visit, how would you rate the services you received?", as we felt it best represented the the patients feelings regarding the true reputation of the centre. First we looked at the distribution of responses, faceted by doctor, to see if any differences can be seen.

The N/A responses represents the patients who don't remember which doctor they saw, or didn't see one (answered something other than Yes to Question 03). For those who remembered which doctor they saw, they tend to exhibit negative skewness in their responses, with the majority of which being in the Good to Excellent range of responses.

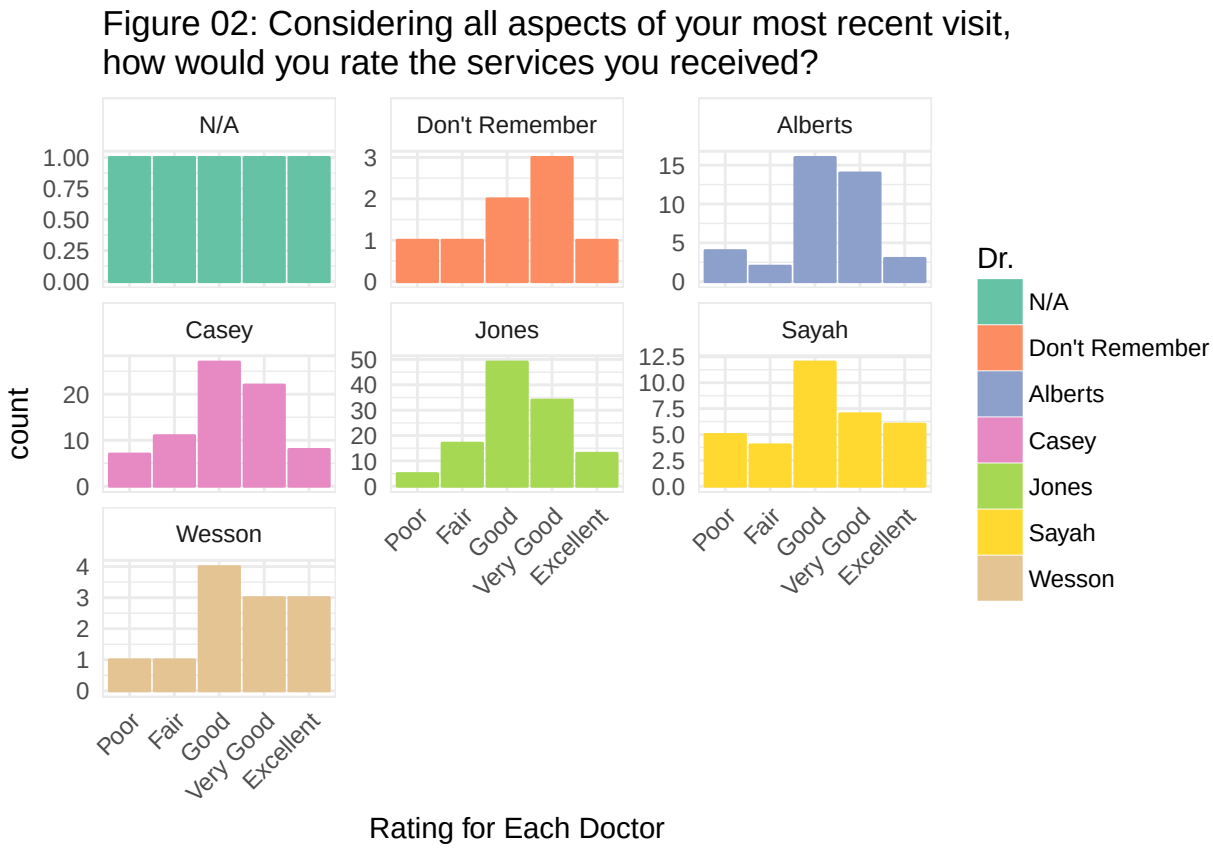


Table 10: Average and Standard Deviation of Rating (Higher is Better)

Dr.	Average Rating	SD of Rating
N/A	3.0000	1.5811
Don't Remember	3.2500	1.2817
Alberts	3.2564	1.0442
Casey	3.1733	1.1074
Jones	3.2797	0.9861
Sayah	3.1471	1.2823
Wesson	3.5000	1.2432

Next we broke down the responses by doctor, and calculated the average rating, with a 5 being Excellent, and 1 being Poor. as well we found the standard deviation of the ratings. From this we found that patients who saw Dr. Wesson gave the highest average rating to the centre, and those who saw Dr. Sayah gave the worst, ignoring the N/A response and Don't Remember.

Table 11: Lower Bound of Wilson Score Confidence Interval

Dr.	Poor	Fair	Good	Very Good	Excellent	Positive	Negative	Number of Patients	Score
N/A	1	1	1	1	1	2.5	2.5	5	0.1704
Don't Remember	1	1	2	3	1	5.0	3.0	8	0.3057
Alberts	4	2	16	14	3	25.0	14.0	39	0.4842
Casey	7	11	27	22	8	43.5	31.5	75	0.4671

Dr.	Poor	Fair	Good	Very Good	Excellent	Positive	Negative	Number of Patients	Score
Jones	5	17	49	34	13	71.5	46.5	118	0.5158
Sayah	5	4	12	7	6	19.0	15.0	34	0.3945
Wesson	1	1	4	3	3	8.0	4.0	12	0.3906

As well, we considered the number of positive responses a patient left, based on which doctor they saw, and the number of negative responses. A response was considered positive if it was Very Good or Excellent, and was considered negative if it was fair or poor. If the response was Good, half a point was assigned to each category. From this we calculated the Wilson Score Confidence Interval, which is given by:

$$\frac{\hat{p} + \frac{z_{\frac{\alpha}{2}}^2}{2n} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p}) + \frac{z_{\frac{\alpha}{2}}^2}{4n}}{n}}}{1 + \frac{z_{\frac{\alpha}{2}}^2}{n}}$$

This interval is a $(1-\alpha)\%$ confidence interval for p, with a center point being the weighted average of $1/2$ and $\hat{p}$. Greater weight is assigned to $\hat{p}$ as the sample size increases. The lower bound is then examined, as it allows us to rank the doctors based on the proportion of positive responses received. From this Dr. Jones is considered the best, with a score of 0.5158, and Dr. Wesson the worst with a score of 0.3906. However, both of them is a noticeable improvement over those who didn't see a doctor, or remember which doctor the may have saw. The difference between these results, and the ones presented previously is that the Wilson Score Interval places a much greater emphasis on sample size compared to merely looking at the averages, and since Dr. Jones saw the most patients (118) and Dr. Wesson seeing the least (12), this may explain the difference in rankings.

Finally a Proportional-Odds Cumulative Logit Model was fitted to the data, as a means of estimating the effect each doctor had on the reputation of the clinic. Let:

$$\pi_i = \mathrm{P}\left(Y = i\right), i = 1, 2, \ldots, J$$

$$L_j = \mathrm{Logit}\left[\mathrm{P}\left(Y \leq j\right)\right] = \ln\left[\frac{\mathrm{P}\left(Y \leq j\right)}{1 - \mathrm{P}\left(Y \leq j\right)}\right] = \ln\left[\frac{\sum_{i=1}^{j}\pi_i}{\sum_{i=j+1}^{J}\pi_i}\right], j = 1, 2, \ldots, J-1$$

In our case we have 5 possible responses, so $J = 5$. We then fit $J - 1$ logistic regression models of the type:

$$L_j = \alpha_j - \left(\sum_{i=1}^{p}\beta_i X_i\right), \forall j = 1, 2, \ldots, J-1$$

Where:

$$L_1 \leq L_2 \leq \cdots \leq L_{J-1}$$

This allows us to keep the β_is constant across the models, and only vary α, which is a monotonically increasing function of j. The higher the β_i terms are, the increased probability (by a multiplicative factor) the response will be in a higher category (i.e. Excellent instead of Poor).

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = `Q10 Response` ~ Dr., data = survey_data_q05)
##
## Coefficients:
##                     Value Std. Error t value
## Dr.Don't Remember 0.5179     1.1283  0.4590
```

```
## Dr.Alberts        0.4770     0.9572  0.4983
## Dr.Casey          0.3030     0.9378  0.3231
## Dr.Jones          0.4281     0.9286  0.4610
## Dr.Sayah          0.2667     0.9709  0.2747
## Dr.Wesson         0.8902     1.0664  0.8348
##
## Intercepts:
##                      Value   Std. Error t value
## Poor|Fair           -2.0176  0.9268    -2.1771
## Fair|Good           -0.9319  0.9170    -1.0163
## Good|Very Good       0.7703  0.9172     0.8399
## Very Good|Excellent  2.3954  0.9271     2.5839
##
## Residual Deviance: 841.8508
## AIC: 861.8508
```

This gives the model (with N/A being our base level):

$$L_j = \alpha_j - (0.5179DR + 0.4770A + 0.3030C + 0.4281J + 0.2667S + 0.8902W)$$

$$\alpha_j = \begin{cases} -2.0176, j = 1 \\ -0.9319, j = 2 \\ 0.7703, j = 3 \\ 2.3954, j = 4 \end{cases}$$

This model believes that Dr. Wesson provides the best rating, and Dr. Sayah the worst. However it is important to note that none of the β_is are statistically significant at $\alpha = 0.05$, which can be confirmed by running an ANOVA of this model against the one containing only the intercept.

Table 12: ANOVA for Cumulative Logit Model

Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	287	843.3523		NA	NA	NA
Dr.	281	841.8508	1 vs 2	6	1.5016	0.9594

This fails to reject the null hypothesis that the two models explain the same amount of variance based on the data, and as such the simpler model should be used.

From this two conclusions can be drawn based on the types of tests being run:

1. Using the Wilson Score, there is a way to rank each of the doctors ability to provide a positive reputation for the centre
2. Using a Cumulative Logistic Regression Model, none of the doctors provide a statistically significant effect towards determine the overall reputation of the centre.

That being said, Dr. Sayah was consistently rated among the lowest of the 5 doctors, and is most likely providing a negative contribution the the centres reputation, and Dr. Wesson should take on more patients to better determine how much he improves the centres reputation. Dr. Jones consistently scored among the highest of the doctors, as well as taking the greatest number of patients.