Homework 5

The exercise is based on the airline customer satisfaction dataset found here: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.

Code book

Variable	Description
Age	Passenger age
Gender	Passenger gender
Type.of.Travel	Purpose of the passenger's flight (personal, business)
Class	Travel class in the plane (business, eco, eco plus)
Customer.Type	Loyal or disloyal customer
Flight.Distance	Flight distance
Inflight.wifi.service	Satisfaction level of inflight wifi service (0: Not applicable; 1-5 where 5 is completely satisfied)
Ease.of.Online.booking	Satisfaction level of online booking
Inflight.service	Satisfaction level of inflight service
Online.boarding	Satisfaction level of online boarding
Inflight.entertainment	Satisfaction level of inflight entertainment
Food.and.drink	Satisfaction level of food and drink
Seat.comfort	Satisfaction level of seat comfort
On.board.service	Satisfaction level of On-board service
Leg.room.service	Satisfaction level of leg room
Departure.Arrival.time.convenient	Satisfaction level of departure/arrival time convenience
Baggage.handling	Satisfaction level of baggage handling
Gate.location	Satisfaction level of gate location
Cleanliness	Satisfaction level of cleanliness

Variable	Description
Checkin.service	Satisfaction level of check-in service
Departure.Delay.in.Minutes	Departure delay in minutes
Arrival.Delay.in.Minutes	Arrival delay in minutes
Satisfaction	Airline satisfaction level (satisfied, neutral, dissatisfied)

An airline called LaneAir is seeking a data scientist consultant to better understand drivers of customer satisfaction. The airline distributed a survey to customers who have flown with LaneAir in the last six months. Customers rated their overall satisfaction as dissatisfied, neutral, or satisfied. Then they rated their satisfaction with various aspects of the flight. The airline also has information on the passengers' flight details, including flight distance and departure delay. LaneAir would like to know which services are worth investing in to improve customer satisfaction. However, they would like the data science consultant to keep in mind that some services are more difficult to improve than others. LaneAir is considering the following investments, ranked from **most difficult** to **least difficult** to implement:

- Newer, larger seats to improve seat comfort and leg room. This will reduce the number of seats per plane, which is LaneAir's last choice
- Newer plane models that improve reliability to minimize delays
- Hire more flight attendants or other staff to improve services including inflight service, cleanliness, or onboarding.
- Technology investment to improve wifi and entertainment service
- Marketing or promotion initiatives to improve customer loyalty or appeal to different customer types (e.g., different age demographic, business/personal travelers)

You will complete this assignment in two parts. The first part of the assignment will be a 3-4 page report that is suitable for **other data scientists.** Here, you will present details of your model to justify the conclusions you presented to the client. This section should present technical details that someone with a data science background can understand. This report must include the following, though you may wish to provide additional details relevant to the analysis:

- Data overview and analysis plan: You should present details of the data that were not included in part 1, for example an appropriate distributional assumption for the outcome variable. Then, present the type of model you used for the analysis. Which type of generalized linear model is best suited for this problem? What is the link function?
- Model results: Present a table of model results including odds ratios, confidence intervals, and p-values. Interpret the results that you think are most compelling.
- Model assessment: Present the confusion matrix and explain your conclusion for the model's predictive accuracy. Additionally, the model you should use for this analysis relies on a key assumption that is unique to this model. Using a different, more "precise"

model, compare predictions using the predictors `Gender` and `Customer type`. Show the confusion matrix for the more precise model and compare the accuracy.
- Conclusion: What do you conclude about the validity of this analysis?

Next, you will create a 2-page report **for the client.** This report should be understood by LaneAir executives with very little understanding of statistics. This report should include the following:

- Introduction: Provide an overview of the dataset and the goals of the analysis. Keep in mind that the LaneAir team is familiar with the questions on the survey, but not the results. So you should include, for example, basic summary statistics to show the team the distribution of customers in the different satisfaction categories.
- Methods: Explain the model you used to analyze the data without getting into technical details. Why did you decide to use that model for this dataset and how does it answer the airline's question?
- Results: What are the key results of the analysis as they relate to the airline's question? Present at least one figure that effectively communicates a key takeaway of the analysis.
- Conclusion: Keeping in mind LaneAir's cost considerations outlined above, what are your recommendations for how they can balance impact with cost? Do you have any recommendations that the airline has not considered? Finally, are there any limitations that the client should be aware of? For example, could certain customers be more likely to respond to the survey than others?

Key

```r
air.sub <- airline[,-1]
air.sub$Satisfaction <- factor(air.sub$Satisfaction)

airmod <- polr(Satisfaction~.,data = air.sub,Hess=TRUE)
summary(airmod)
```

```
Call:
polr(formula = Satisfaction ~ ., data = air.sub, Hess = TRUE)

Coefficients:
                                       Value Std. Error   t value
GenderMale                        -5.763e-03  7.438e-02  -0.07748
Customer.TypeLoyal Customer        1.657e+00  1.136e-01  14.58061
Age                               -5.128e-03  2.423e-03  -2.11638
Type.of.TravelPersonal Travel     -1.990e+00  1.091e-01 -18.23193
ClassEco                          -3.530e-01  8.774e-02  -4.02312
ClassEco Plus                     -6.286e-01  1.394e-01  -4.50913
Flight.Distance                    1.666e-05  5.063e-05   0.32902
Inflight.wifi.service              3.149e-01  4.639e-02   6.78853
Departure.Arrival.time.convenient -4.461e-02  2.928e-02  -1.52341
Ease.of.Online.booking            -2.914e-01  4.471e-02  -6.51819
Gate.location                     -8.317e-03  3.316e-02  -0.25079
Food.and.drink                    -2.274e-02  4.215e-02  -0.53939
Online.boarding                    4.135e-01  3.898e-02  10.60664
Seat.comfort                       5.510e-02  4.213e-02   1.30781
Inflight.entertainment             5.434e-02  5.242e-02   1.03671
On.board.service                   1.529e-01  3.603e-02   4.24400
Leg.room.service                   1.646e-01  3.023e-02   5.44319
Baggage.handling                   7.587e-02  4.173e-02   1.81797
Checkin.service                    1.901e-01  3.146e-02   6.04382
Inflight.service                   1.249e-01  4.203e-02   2.97131
Cleanliness                        1.113e-01  4.872e-02   2.28536
Departure.Delay.in.Minutes         4.690e-03  3.659e-03   1.28196
Arrival.Delay.in.Minutes          -7.395e-03  3.629e-03  -2.03787

Intercepts:
                     Value    Std. Error t value
dissatisfied|neutral   3.0334   0.0466    65.0710
neutral|satisfied      4.8350   0.0711    67.9927
```

```
Residual Deviance: 5466.347
AIC: 5516.347
```

```
pvals <- pnorm(-abs(summary(airmod)$coef[,"t value"]))*2

cbind(OR=exp(summary(airmod)$coefficients[,1]),exp(confint(airmod)),pvals)
```

```
Waiting for profiling to be done...

Warning in cbind(OR = exp(summary(airmod)$coefficients[, 1]),
exp(confint(airmod)), : number of rows of result is not a multiple of vector
length (arg 1)

                                         OR     2.5 %    97.5 %        pvals
GenderMale                        0.9942538 0.8587915 1.1511279 9.382416e-01
Customer.TypeLoyal Customer       5.2420774 4.1691637 6.6018506 3.731430e-48
Age                               0.9948851 0.9899747 0.9998096 3.431287e-02
Type.of.TravelPersonal Travel     0.1367068 0.1094975 0.1702741 2.879582e-74
ClassEco                          0.7025717 0.5751105 0.8584993 5.743325e-05
ClassEco Plus                     0.5333282 0.3964524 0.7171900 6.509267e-06
Flight.Distance                   1.0000167 0.9999240 1.0001096 7.421430e-01
Inflight.wifi.service             1.3701305 1.2512877 1.5011290 1.132780e-11
Departure.Arrival.time.convenient 0.9563695 0.9028653 1.0130119 1.276568e-01
Ease.of.Online.booking            0.7471945 0.6841341 0.8155055 7.116070e-11
Gate.location                     0.9917172 0.9269949 1.0610184 8.019784e-01
Food.and.drink                    0.9775210 0.8973521 1.0643831 5.896156e-01
Online.boarding                   1.5120726 1.4006852 1.6333727 2.775401e-26
Seat.comfort                      1.0566423 0.9722759 1.1481540 1.909387e-01
Inflight.entertainment            1.0558430 0.9491974 1.1744661 2.998712e-01
On.board.service                  1.1652347 1.0849689 1.2515506 2.195748e-05
Leg.room.service                  1.1788904 1.1103566 1.2517197 5.233557e-08
Baggage.handling                  1.0788194 0.9924243 1.1727858 6.906942e-02
Checkin.service                   1.2094025 1.1359155 1.2879140 1.505052e-09
Inflight.service                  1.1330021 1.0410701 1.2330984 2.965359e-03
Cleanliness                       1.1177737 1.0154006 1.2302589 2.229199e-02
Departure.Delay.in.Minutes        1.0047014 0.9975379 1.0119606 1.998581e-01
Arrival.Delay.in.Minutes          0.9926320 0.9855630 0.9997021 4.156299e-02
```

Students should use the ordinal regression model for this problem. Students can make different arguments based on the proposed investments from the airline. Looking at the t-statistics, the biggest drivers of customer satisfaction are type of travel, type of customer, online boarding, online booking, wifi service, and inflight service. A marketing/promotion initiative to gain more loyal customers strikes a balance between client preference and model output. But students may have different answers that are reasonable. Some students may exclude variables because the airline cannot intervene (e.g., gate location). Model selection is not necessary here, particularly with the large sample size.

Students should check the **proportional odds assumption** by comparing predictions from the proportional odds model and the multinomial logistic model. They should create a new dataset that includes different values for `gender` and `customer type`, holding all other variables to a single value. Exact implementation may vary here.

```
confusionMatrix(predict(airmod),air.sub$Satisfaction)
```

```
Confusion Matrix and Statistics

              Reference
Prediction     dissatisfied neutral satisfied
  dissatisfied          502     511        76
  neutral               300     277       114
  satisfied             186     144      1368

Overall Statistics

               Accuracy : 0.6173
                 95% CI : (0.6009, 0.6335)
    No Information Rate : 0.448
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4012

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: dissatisfied Class: neutral Class: satisfied
Sensitivity                       0.5081        0.29721           0.8780
Specificity                       0.7643        0.83739           0.8281
Pos Pred Value                    0.4610        0.40087           0.8057
Neg Pred Value                    0.7966        0.76498           0.8933
Prevalence                        0.2841        0.26797           0.4480
Detection Rate                    0.1443        0.07964           0.3933
Detection Prevalence              0.3131        0.19868           0.4882
```

```
Balanced Accuracy                 0.6362        0.56730           0.8531
```

```
multmod <- multinom(Satisfaction~.,data=air.sub)
```

```
# weights:  75 (48 variable)
initial  value 3820.973540
iter  10 value 3188.373372
iter  20 value 2850.280130
iter  30 value 2696.830475
iter  40 value 2532.514648
iter  50 value 2471.655249
final  value 2455.334729
converged
```

```
newdata <- data.frame(Gender=c("Female","Male","Male","Female"),
                      Customer.Type=c("Loyal Customer","Loyal Customer",
                                      "disloyal Customer","disloyal Customer"),
                      Age=mean(air.sub$Age),
                      Type.of.Travel=c("Personal Travel","Personal Travel",
                                       "Personal Travel","Personal Travel"),
                      Class=c("Eco","Eco","Eco","Eco"),
                      Flight.Distance=mean(air.sub$Flight.Distance),
                      Inflight.wifi.service=mean(air.sub$Inflight.wifi.service),
                      Departure.Arrival.time.convenient=
                        mean(air.sub$Departure.Arrival.time.convenient),
                      Ease.of.Online.booking=mean(air.sub$Ease.of.Online.booking),
                      Gate.location=mean(air.sub$Gate.location),
                      Food.and.drink=mean(air.sub$Food.and.drink),
                      Online.boarding=mean(air.sub$Online.boarding),
                      Seat.comfort=mean(air.sub$Seat.comfort),
                      Inflight.entertainment=mean(air.sub$Inflight.entertainment),
                      On.board.service=mean(air.sub$Inflight.entertainment),
                      Leg.room.service=mean(air.sub$Leg.room.service),
                      Baggage.handling=mean(air.sub$Baggage.handling),
                      Checkin.service=mean(air.sub$Baggage.handling),
                      Inflight.service=mean(air.sub$Inflight.service),
                      Cleanliness=mean(air.sub$Cleanliness),
                      Departure.Delay.in.Minutes=mean(air.sub$Departure.Delay.in.Minutes),
                      Arrival.Delay.in.Minutes=mean(air.sub$Arrival.Delay.in.Minutes))

predict(airmod,newdata,type="probs")
```

```
  dissatisfied   neutral  satisfied
1    0.3877775 0.4055221 0.20670039
2    0.3891465 0.4050964 0.20575703
3    0.7695579 0.1833499 0.04709218
4    0.7685344 0.1841142 0.04735146
```

```
predict(multmod,newdata,type="probs")
```

```
  dissatisfied   neutral  satisfied
1    0.4627670 0.4446957 0.09253727
2    0.4535798 0.4629508 0.08346943
3    0.5190939 0.4701031 0.01080305
4    0.5332606 0.4546801 0.01205926
```

```
confusionMatrix(predict(multmod),air.sub$Satisfaction)
```

```
Confusion Matrix and Statistics

              Reference
Prediction     dissatisfied neutral satisfied
  dissatisfied          464     423       116
  neutral               341     368        61
  satisfied             183     141      1381

Overall Statistics

               Accuracy : 0.6363
                 95% CI : (0.62, 0.6523)
    No Information Rate : 0.448
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4309

 Mcnemar's Test P-Value : 5.377e-12

Statistics by Class:

                     Class: dissatisfied Class: neutral Class: satisfied
Sensitivity                       0.4696         0.3948           0.8864
Specificity                       0.7835         0.8421           0.8313
Pos Pred Value                    0.4626         0.4779           0.8100
```

```
Neg Pred Value                    0.7883         0.7917           0.9002
Prevalence                        0.2841         0.2680           0.4480
Detection Rate                    0.1334         0.1058           0.3971
Detection Prevalence              0.2884         0.2214           0.4902
Balanced Accuracy                 0.6266         0.6185           0.8588
```

Answers may vary, but it seems we do have evidence that the proportional odds assumption may be violated based on the predicted probabilities. We see a slight improvement in predictive accuracy using the multinomial logistic model over the ordinal model, but probably not enough to sacrifice the better interpretability of the ordinal model. In both models, the sensitivity is pretty low for the dissatisfied and neutral levels.