MATH 324 Final Project

Nathan Howland, Madilyn Webb, Ellise Putnam, Addison Hart

1. Project Description

This experiment is using a data set of the weather patterns across various cities in Australia, which contains variables such as the humidity, pressure, cloud coverage, temperature, and much more. This experiment was conducted as an observational study, where the goal of the study was to predict the possibility of rain based on these variables. We wanted to determine: 1) Whether additional weather variables led to a sufficiently improved model for predicting rainfall in Australia 2) Whether there are different variables that conclude better predictions based on the city within Australia

1.1 Research Questions

Research Question 1: What model should we use to predict rainfall across Australia? We want to discover which variables are the most reliable predictors of the rainfall for the "Next day".

Research Question 2: Should we use different models for different locations? Are there models that are better suited for different locations across Australia, and what variables are the best for these locations?

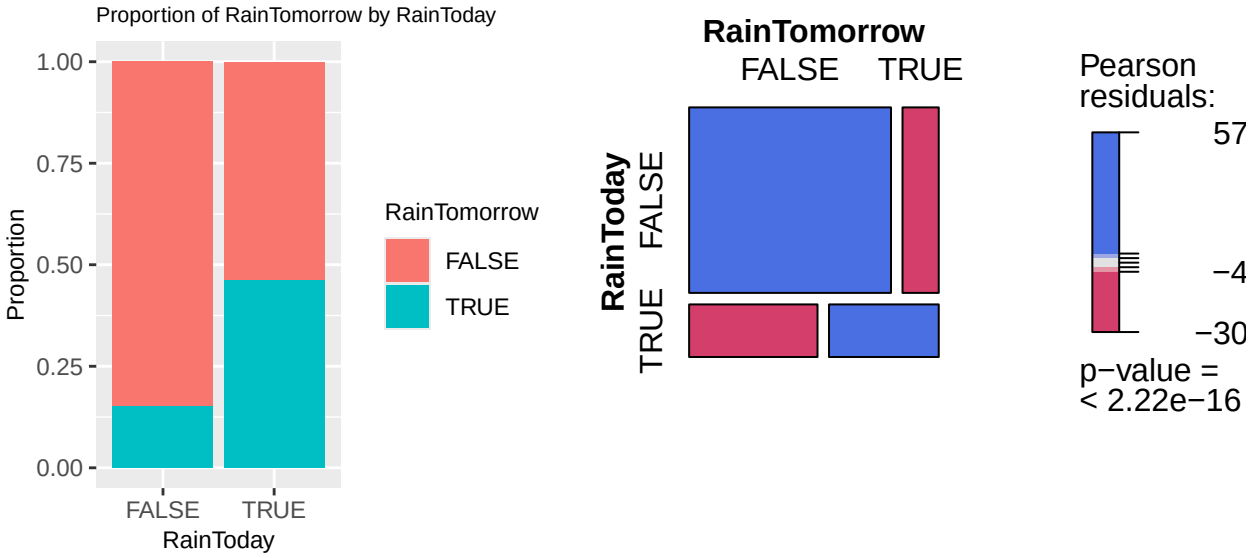
1.2 Variables

Variable	Type	E/R
Location	Categorical	E
MinTemp	Quantitative	E
MaxTemp	Quantitative	E
Rainfall	Quantitative	E
Evaporation	Quantitative	E
Sunshine	Quantitative	E
WindGustDir	Categorical	E
WindGustSpeed	Quantitative	E
WindDir9am	Categorical	E
WindDir3pm	Categorical	E
WindSpeed9am	Quantitative	E
WindSpeed3pm	Quantitative	E
Humidity9am	Quantitative	E
Humidity3pm	Quantitative	E
Pressure9am	Quantitative	E
Pressure3pm	Quantitative	E
Cloud9am	Categorical	E
Cloud3pm	Categorical	E
Temp9am	Quantitative	E
Temp3pm	Quantitative	E
RainToday	Boolean	E

Variable	Type	E/R
RainTomorrow	Boolean	R

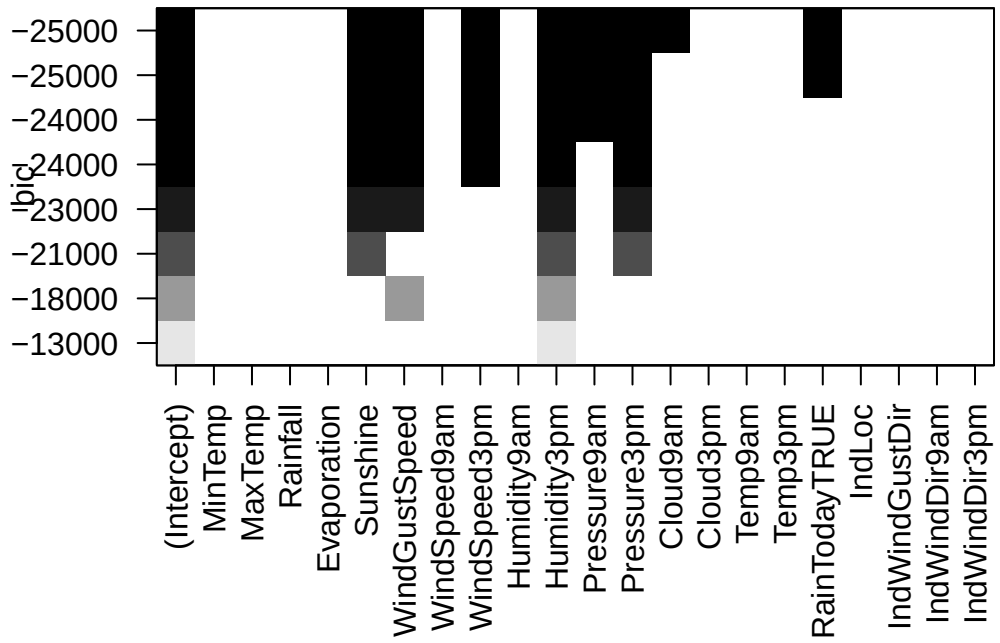
2. Detailed Exploratory Data Analysis (EDA)

Question 1: Simple logistic regression model using RainToday as a predictor



For this initial analysis, we are going to use a cleaned dataset that only contains RainToday and RainTomorrow. Based on the plots above, we can see that there is a significant difference in the proportion of rain tomorrow based on whether it rained today. The mosaic plot shows that the proportion of rain tomorrow is higher when it rained today compared to when it didn't. This initial analysis suggests that RainToday is a pretty solid predictor of RainTomorrow.

Question 2: Best model (using BIC) of all locations

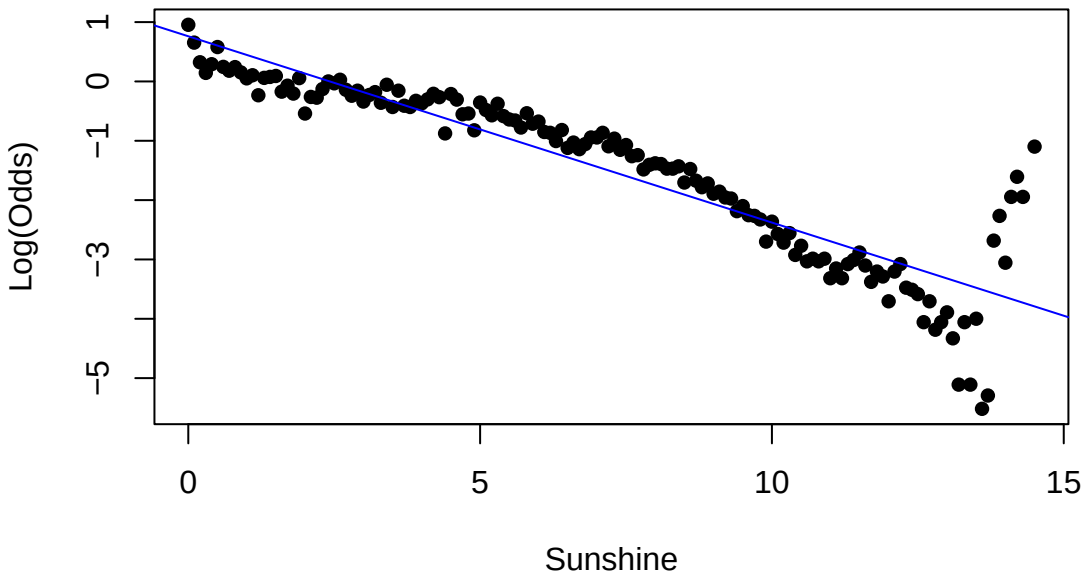


After performing a stepwise function for all variables, it was concluded from solely the BIC table that the most effective variables for predicting RainTomorrow was Sunshine + WindGustSpeed + WindSpeed3pm + Humidity3pm + Pressure3pm + RainToday. Additionally, when creating the BIC table, some additional cleanup was performed where we created Indicator varaibles for some of the categorical variables such as IndLoc, IndWindGustDir, IndWindDir9am and IndWindDir3pm. Using our BIC table we created a model using these variables, as seen below.

As demonstrated from the following model, all variables prove to be significant predictors of RainTomorrow as all p-values are less than our significance level of 0.05. Additionally, we have an AIC value of 38,540 which is the lowest value we could attain for predicting RainTomorrow. For some context, when all other predictors are zero, the log odds of RainTomorrow is 74.50. Additionally, for each unit increase in sunshine, the log odds of RainTomorrow decrease by 0.168. This interpretation of variables also apply meaning that, for example, if it rained today (RainTodayTRUE), the log odds of RainTomorrow increase by 0.389 compared to when it didn't rain today.

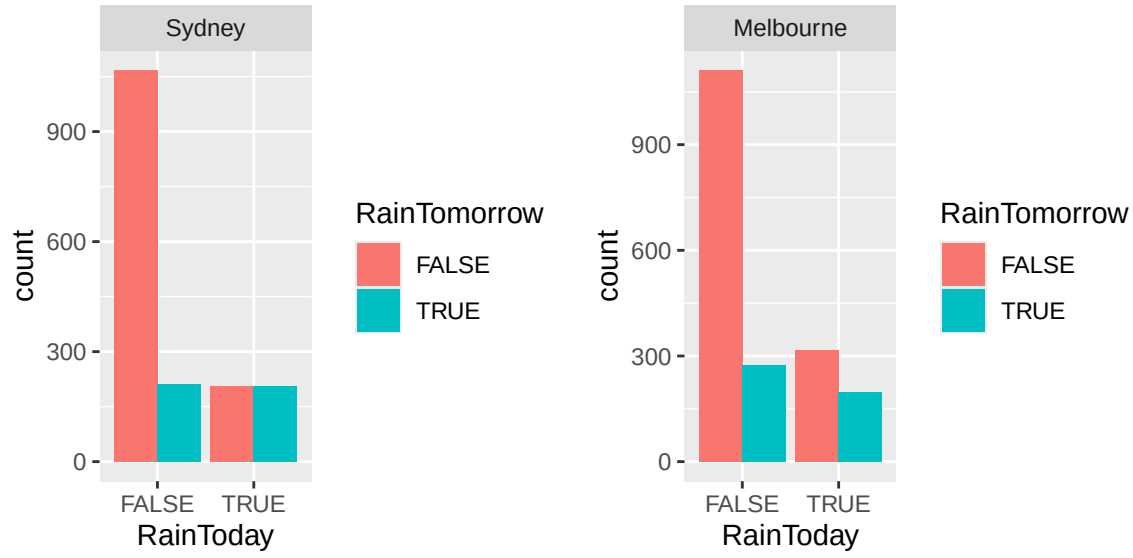
$$\hat{\text{RainTomorrow}} = 74.4980470 + -0.1679587 * \text{Sunshine} + 0.0508944 * \text{WindGustSpeed} - 0.0330498 *$$

$$\text{WindSpeed3pm} + 0.0533006 * \text{Humidity3pm} - 0.0781103 * \text{Pressure3pm} + 0.3888131 * \text{RainTodayTRUE}$$



As a result of our predictive model, we created an empirical logit plot to verify whether the relationships between Sunshine + WindGustSpeed + WindSpeed3pm + Humidity3pm + Pressure3pm + RainToday and RainTomorrow align with our expectations to assess the overall goodness of fit of your logistic regression model. Given the plot demonstrates data points that reasonably fall along the blue line, we can conclude that our model effectively demonstrates the relationship between our predictors and the RainTomorrow.

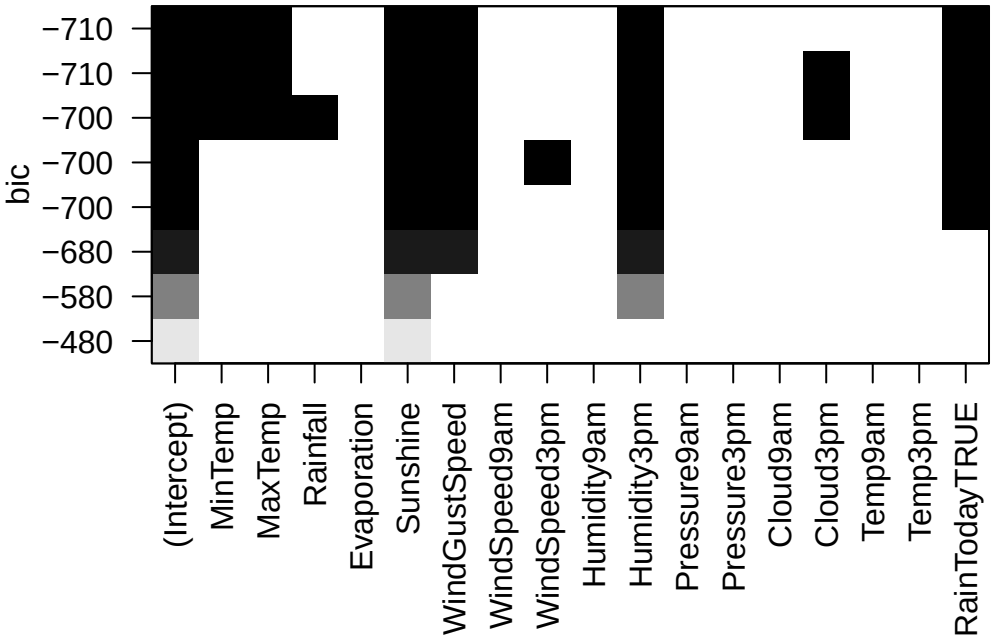
Question 3: Simple Logistic regression model dependent on the location

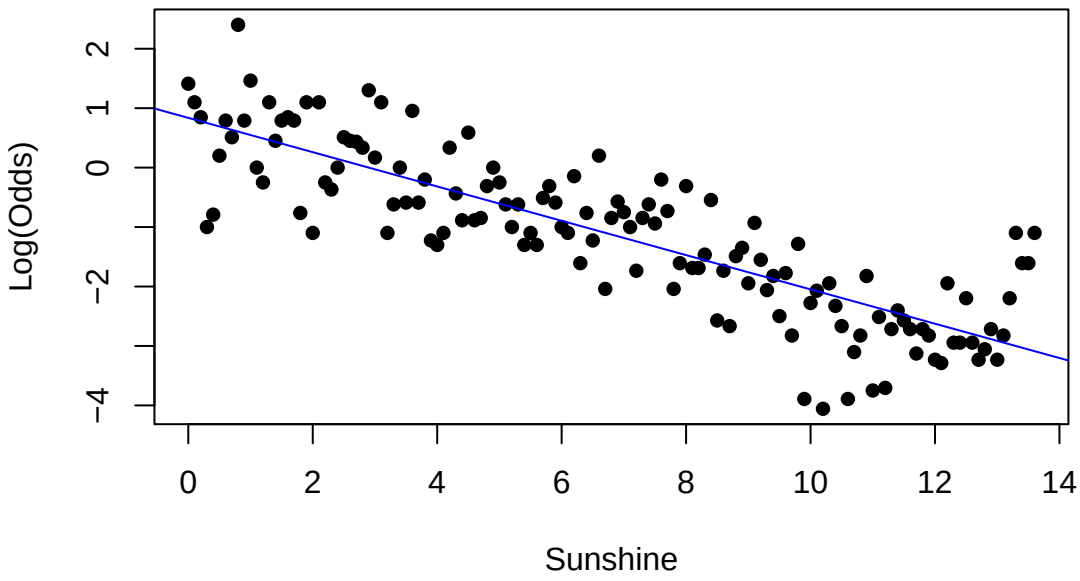


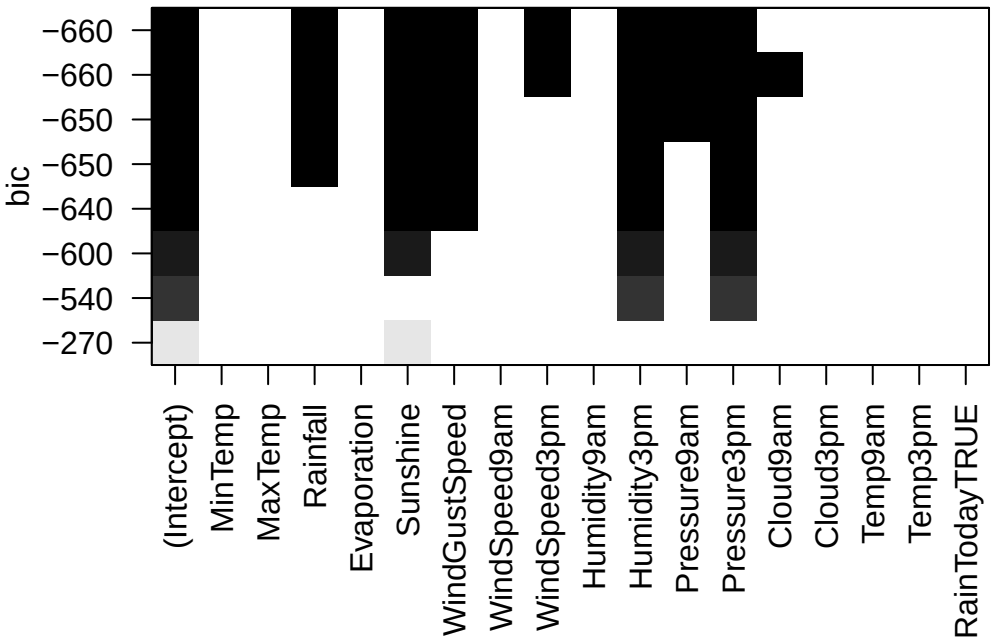
The patterns observed in the box plots above show that there is no outstanding initial patterns to be detected between the two locations. This suggests the need for further investigation / analysis of the dataset

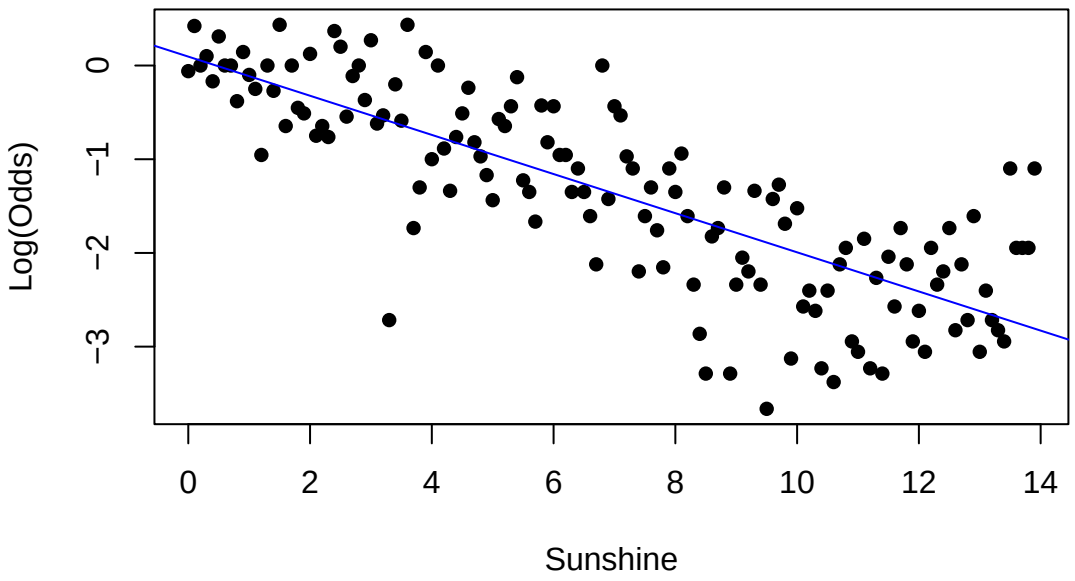
to determine if there is a model that is bette suited for different locations.

Question 4: Best model (using BIC) of two locations




Log(Odds)
2
1
0
−2
−4
0
2
4
6
8
10
12
14
Sunshine


bic
−660
−660
−650
−650
−640
−600
−540
−270
(Intercept)
MinTemp
MaxTemp
Rainfall
Evaporation
Sunshine
WindGustSpeed
WindSpeed9am
WindSpeed3pm
Humidity9am
Humidity3pm
Pressure9am
Pressure3pm
Cloud9am
Cloud3pm
Temp9am
Temp3pm
RainTodayTRUE



A BIC plot is made of both data from Sydney and data from Melbourne to derive the explanatory variables that maximize the BIC score, which we then use to create a simple logistic model. A linearity condition assesment using the logarithm of the odds is also employed.

3. Statistical Analysis

Simple logistic regression model using RainToday as a predictor

Simple Logistic regression model dependent on the location

The model assumptions we need to satisfy for these two models are:

- Linearity: We cannot prove this assumption because a model using only one categorical boolean variable is not a line. This model essentially uses the number of datapoints having rain yesterday in order to determine a value for rain today.
- Independence: The datapoints were obtained from a sensor whose results are not dependent on yesterday's results, so they are independent.
- Randomness: The datapoints are always recorded from a sensor located in the same geographical location. For example, datapoints from the Sydney Airport are always recorded near the airport. That means that there is no randomness of recordings in the general area indicated by the datapoint. We cannot, therefore, generalize to the entire city.

Best model (using BIC) of all locations

Best model (using BIC) of two locations and the best model (using BIC) of all locations

The model assumptions we need to satisfy for these two models are:

- Linearity: We satisfied this requirement using an Empirical Logit plot, which tests linearity for categorical variables. The best model for all locations is somewhat convex, which could suggest this assumption is not true. The best model for two locations is completely linear, however.
- Independence: The datapoints were obtained from a sensor whose results are not dependent on yesterday's results, so they are independent.
- Randomness: The datapoints are always recorded from a sensor located in the same geographical location. For example, datapoints from the Sydney Airport are always recorded near the airport. That means that there is no randomness of recordings in the general area indicated by the datapoint. We cannot, therefore, generalize to the entire city.

Research Question 1

Question 1: Simple logistic regression model using RainToday as a predictor

After the statistical analysis of this simple logistic regression model, we found that it is fairly sufficient at predicting RainTomorrow. This model obtained a Brier score of 0.165, which is a relatively low score, which shows that the model is making fairly accurate predictions of rain tomorrow. The odds ratio of 4.84 indicates that the odds of rain tomorrow are 4.84 times higher when it rained today compared to when it didn't rain today.

Question 2: Best model (using BIC) of all locations

Given our brier score of 0.1066608 which indicates that, on average, the squared difference between the predicted probabilities of rain tomorrow and the actual outcomes is 0.1066608. Overall, our score of 0.1066608 indicates that our model is making accurate predictions of rain tomorrow, which displays a greater accuracy than the simple logistic regression model from question 1.

Research Question 2

Question 3: Simple Logistic regression model dependent on the location

In the basic forecast model where we solely used RainToday to predict RainTomorrow, we calculated the brier score and odds ratio for both Sydney and Melbourne. The brier scores for both Sydney and Melbourne are very similar with values of 0.165 and 0.179 respectively. Once we create a model with more explanatory variables such as Sunshine, MinTemp, MaxTemp, WindGustSpeed, Humidity3pm, and RainToday, we get much lower Brier Scores for both Sydney and Melbourne with values of 0.118 and 0.137, which indicates that this model has greater accuracy in predicting RainTomorrow.

Question 4: Best model (using BIC) of two locations

Need Evaluation of Question 4 Results

4. Conclusions

Research Question 1: Based on the results of our analysis, and a Brier score of 0.107, we can conclude that the best model for predicting rainfall across Australia is the model that includes the variables Sunshine, WindGustSpeed, WindSpeed3pm, Humidity3pm, Pressure3pm, and RainToday. However, we can also conclude that the simple logistic regression model using RainToday as a predictor is also a reliable model for predicting rainfall across Australia, especially when without access to the extra variables used in the best model.

Research Question 2: From the results of this experiment, we obtained Brier Scores for the simple logistic regression and the best BIC models for both Sydney and Melbourne (0.165 vs. 0.179, 0.118 vs 0.137). Running a paired t-test on these Brier scores, we obtained a p-value of 0.096, which concludes there is not enough evidence that there is no significanct difference between the Brier scores of the different locations. Therefore, we can conclude that a model used for predicting rainfall in Sydney would yield similar results to a model used for predicting rainfall in Melbourne.

5. Appendix

Question 1 Code

```
# clean data
clean_weather <- weather %>% na.omit()

# model
lr_model <- glm(RainTomorrow ~ RainToday, family = binomial(link = "logit"), data = clean_weather)

# eda
ggplot(clean_weather, aes(x = RainToday, fill = RainTomorrow)) +
  geom_bar(position = "fill") +
  ylab("Proportion") +
  ggtitle("Proportion of Rain Tomorrow by Rain Today")

mosaic(~ RainToday + RainTomorrow, data = clean_weather, shade = TRUE, legend = TRUE)

emplogitplot1(as.numeric(RainTomorrow) ~ as.numeric(RainToday), data = clean_weather, ngroups = "all")

# Brier Score
prediction <- predict(lr_model, clean_weather, type = "response")
BriarScore <- mean((prediction - clean_weather$RainTomorrow)^2)

# Odds Ratio
pi_rainToday <- exp(-1.720027 + 1.576018) / (1 + exp(-1.720027 + 1.576018))
odds_rainToday <- pi_rainToday / (1 - pi_rainToday)
pi_not_rainToday <- exp(-1.720027) / (1 + exp(-1.720027))
odds_not_rainToday <- pi_not_rainToday / (1 - pi_not_rainToday)
```

Question 2 Code

```r
# clean data
clean_weather2 <- clean_weather
clean_weather2$Location <- factor(clean_weather2$Location)

# adding indicators
clean_weather$IndLoc <- ifelse(clean_weather$Location == "Sydney", 1, 0)
clean_weather$IndWindGustDir <- ifelse(clean_weather$WindGustDir == "W", 1, 0)
clean_weather$IndWindDir9am <- ifelse(clean_weather$WindDir9am == "W", 1, 0)
clean_weather$IndWindDir3pm <- ifelse(clean_weather$WindDir3pm == "WNW", 1, 0)

# performing stepwise and creating BIC
none <- lm(RainTomorrow ~ 1, data = clean_weather)
full_model <- lm(RainTomorrow ~ ., data = clean_weather)

step(none, scope = list(upper = full_model), direction = "forward", really.big = T)

all <- regsubsets(RainTomorrow ~ . - Location - WindGustDir - WindDir9am - WindDir3pm, data = clean_wea
plot(all, scale = "bic")

summaryHH(all)

# create model based on BIC
myModel <- glm(RainTomorrow ~ Sunshine + WindGustSpeed + WindSpeed3pm + Humidity3pm + Pressure3pm + Rai
summary(myModel)

# plot model
emplogitplot1(RainTomorrow ~ Sunshine + WindGustSpeed + WindSpeed3pm + Humidity3pm + Pressure3pm + Rain

predicted <- predict(myModel, newdata = clean_weather, type = "response")


brier_score <- compute_brier(clean_weather$RainTomorrow, predicted)
brier_score
```

Question 3 Code

```r
cleanWeather3 <- weather %>% na.omit()
SydneyWeather <- cleanWeather3 %>% filter(Location == "Sydney")
MelbourneWeather <- cleanWeather3 %>% filter(Location == "Melbourne")

gf_bar(~ RainToday | Location, data = SydneyWeather, fill = ~RainTomorrow, position = position_dodge())
gf_bar(~ RainToday | Location, data = MelbourneWeather, fill = ~RainTomorrow, position = position_dodge

model_sydney <- glm(RainTomorrow ~ RainToday, family = binomial(link = "logit"), data = SydneyWeather)
model_melbourne <- glm(RainTomorrow ~ RainToday, family = binomial(link = "logit"), data = MelbourneWea
summary(model_sydney)
summary(model_melbourne)

plot(model_sydney, which = 1)
plot(model_melbourne, which = 1)
emplogitplot1(as.numeric(RainTomorrow) ~ as.numeric(RainToday), data = SydneyWeather, ngroups = "all")
emplogitplot1(as.numeric(RainTomorrow) ~ as.numeric(RainToday), data = MelbourneWeather, ngroups = "all
```

```r
# Make predictions for Sydney
predictions_sydney <- predict(model_sydney, newdata = SydneyWeather, type = "response")

# Make predictions for Melbourne
predictions_melbourne <- predict(model_melbourne, newdata = MelbourneWeather, type = "response")

# Compute Brier scores for Sydney
brier_score_sydney <- compute_brier(SydneyWeather$RainTomorrow, predictions_sydney)
brier_score_sydney

# Compute Brier scores for Melbourne
brier_score_melbourne <- compute_brier(MelbourneWeather$RainTomorrow, predictions_melbourne)
brier_score_melbourne

# Calculate odds ratios
pi_sydney <- exp(-1.58290 + 1.62910) / (1 + exp(-1.58290 + 1.62910))
odds_sydney <- pi_sydney / (1 - pi_sydney)
odds_sydney

pi_melbourne <- exp(-1.50550 + 1.02034) / (1 + exp(-1.50550 + 1.02034))
odds_melbourne <- pi_melbourne / (1 - pi_melbourne)
odds_melbourne
```

Question 4 Code

```r
# Data cleaning
weather_cleaned_sydney <- weather %>%
  dplyr::select(-c(WindGustDir, WindDir9am, WindDir3pm)) %>%
  na.omit() %>%
  filter(Location == "Sydney") %>%
  dplyr::select(-c(Location))
weather_cleaned_melbourne <- weather %>%
  dplyr::select(-c(WindGustDir, WindDir9am, WindDir3pm)) %>%
  na.omit() %>%
  filter(Location == "Melbourne") %>%
  dplyr::select(-c(Location))

# BIC plot Sydney
all_sydney <- regsubsets(RainTomorrow ~ ., data = weather_cleaned_sydney, method = "exhaustive")
plot(all_sydney, scale = "bic")
summaryHH(all_sydney)
rain_tomorrow_sydney_model <- glm(RainTomorrow ~ MinTemp + MaxTemp + Sunshine + WindGustSpeed + Humidity
# Linearity odds plot Sydney
emplogitplot1(RainTomorrow ~ Sunshine + MinTemp + MaxTemp + WindGustSpeed + Humidity3pm + RainToday, da

# cor(weather_cleaned_sydney)

# BIC plot Melbourne
all_melbourne <- regsubsets(RainTomorrow ~ ., data = weather_cleaned_melbourne, method = "exhaustive")
plot(all_melbourne, scale = "bic")
summaryHH(all_melbourne)
rain_tomorrow_melbourne_model <- glm(RainTomorrow ~ MinTemp + MaxTemp + Sunshine + WindGustSpeed + Humi
```

```r
# Linearity odds plot Melbourne
emplogitplot1(RainTomorrow ~ Sunshine + MinTemp + MaxTemp + WindGustSpeed + Humidity3pm + RainToday, da

# cor(weather_cleaned_melbourne)

# Brier scores
print(paste("Brier score for Sydney:", compute_brier(weather_cleaned_sydney$RainTomorrow, predict(rain_
print(paste("Brier score for Melbourne:", compute_brier(weather_cleaned_melbourne$RainTomorrow, predict
```