Final Project Report

Group E8

05/04/2022

Introduction

Public food safety is of concern to the general population. The Food Hygiene Agency in the United Kingdom puts out a report on the food hygiene rating in the boroughs of London from 2002 to 2022. We will use the UK Food Hygiene Rating Data (London). The data is collected by multiple participating local authorities in the London area and was aggregated by the Food Hygiene Agency for honest communication between the government and citizens with topic of Health in mind, and published by DATOTA (2022) on Kaggle. Based on both governmental and our concern, we will investigate the relationship between the food hygiene evaluation and the date of inspection, area (health authority), and business type of the food service provider. In this study, we want to generate a model that causally explains what determines good or bad in the hygiene evaluation of restaurants across London and, possibly, make predictive assessments such that it can help the public have more tooling and information for assessing the outcome of future restaurants hygiene inspections to help empower consumer consciousness. In the dataset, we will define a new variable food hygiene rating with a rating of 3 to 5 as 'good' hygiene condition and 0 to 2 as 'not good' hygiene condition in order to make our food hygiene evaluation a binary response variable. Furthermore, we will exclude all the businesses that have not gotten reviewed or have not been reviewed yet. For our model, there are 3 explanatory variables in our study. The food business type is a categorical variable with 5 levels: Restaurant/Cafe/Canteen, Retailers-other, Takeaway/sandwich shop, Other catering premises, and Hospitals/Childcare/Caring Premises. The date of inspection is the date at which the hygiene review took place, from 2002 to 2022, and we will exclude the data that does not have a date.

Analysis

In order to study the proposed relationship (the casual analysis), the first steps are tidying and wrangling the data, and then performing descriptive data analysis. For tidying, the missing values were checked for and removed, the `RatingValue` has two factors, businesses that did not get reviewed and businesses that have yet to be reviewed, which will be removed, and extraneous variables were dropped such that the data set was well-structured and relevant to the proposed relationship. For wrangling, the `RatingDate` variable was split into three separate variables (`Year`, `Month`, and `Day`), the variable `RatingValue` was transformed into the new response variable `Pass` which is a factor of levels 0 and 1 where 0 represents the rating below 3 and 1 represents the rating value equal to or larger than 3 (i.e. converting ordinal factors into nominal bins).

For the descriptive analysis, several plots of the exploratory variables and the response variable were created. Since the response variable `Pass` is categorical, we agreed that bar plots and histograms are most suited to explore the data. The plots below include one histogram of `LocalAuthorityCode`, and two bar plots of `BusinessType` and `Pass`:

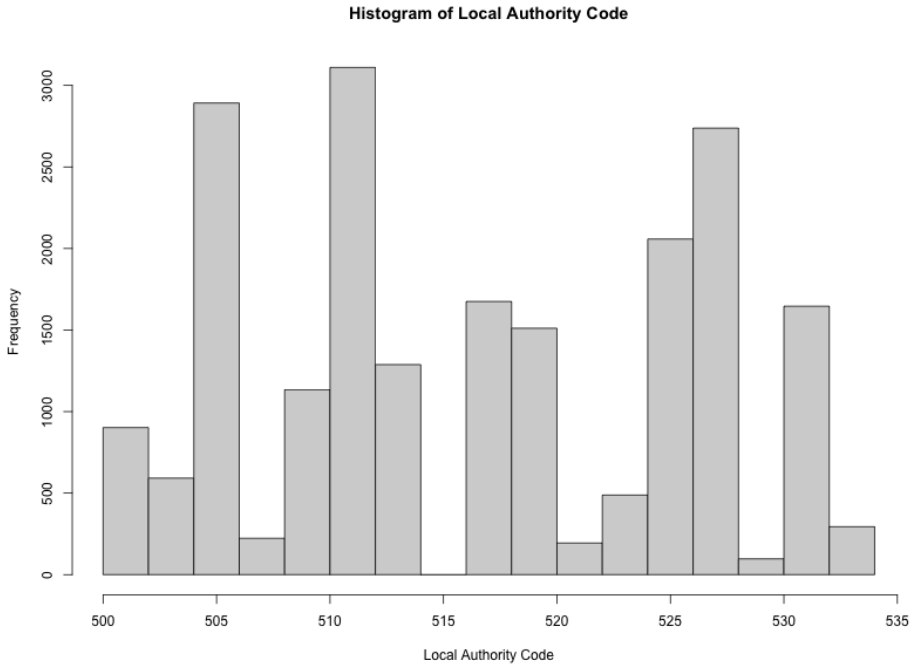

Figure 1: Histogram of Local Authority Code

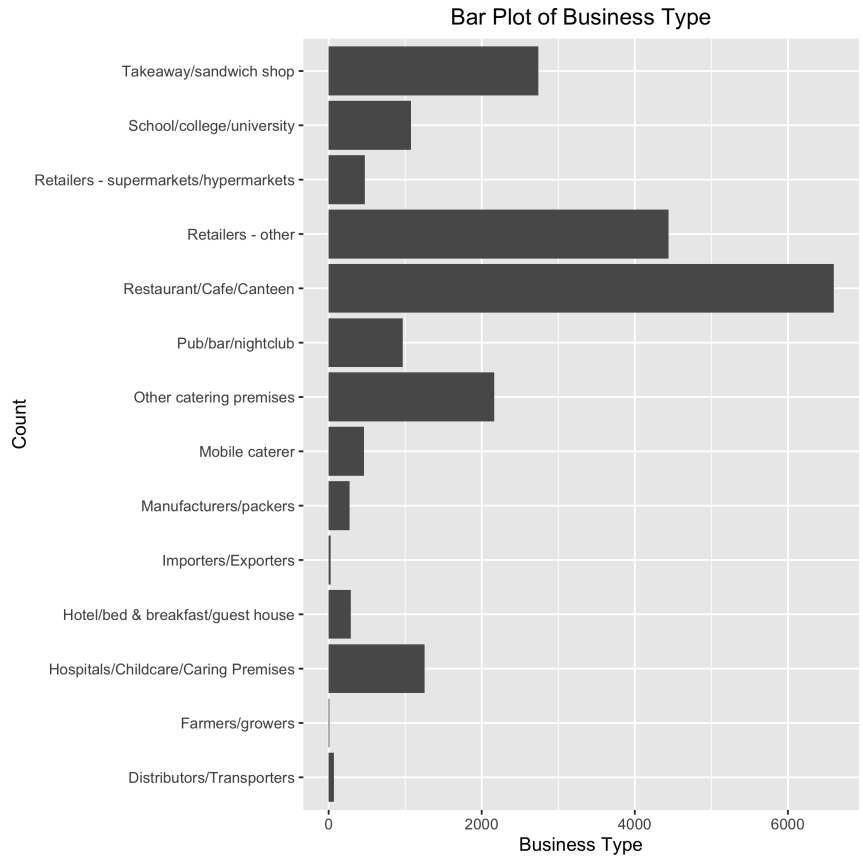

Figure 2: Bar Plot of Business Type

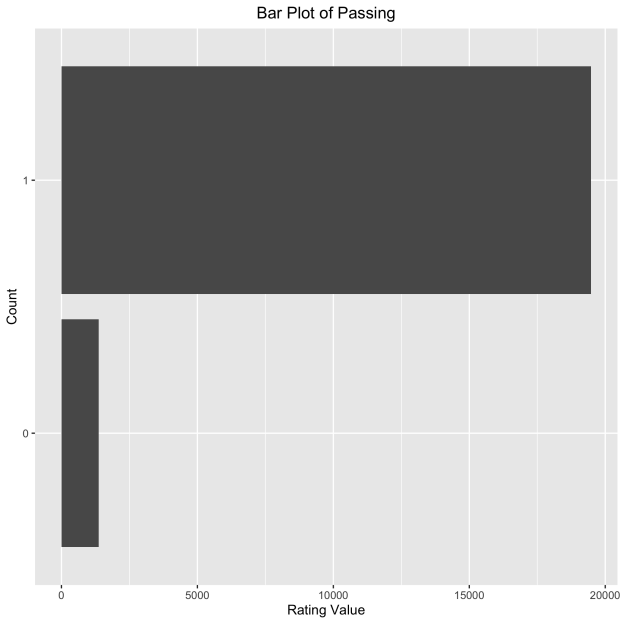

Figure 3: Bar Plot of Rating Value

We used the train-test methodology to evaluate the final model in this project. As such, it is essential to split the data into two data-sets: training and testing. We used 75% of the data to fit the model, and 25% of the data to test the prediction accuracy.

Since there are not that many variables, we decided to train a full logistic regression model containing all of the variables. Among the propose explanatory variables, there is one variable that is treated as a factor, `BusinessType`. Furthermore, the `LocalAuthorityCode` variable is comprised of discrete numbers which allows it to be treated as a factor as well; however, the values variate and it was deemed to overly increase the complexity of the model with marginal gain and, thus, was kept as a numeric variable. Moreover, there are no interaction terms in the model as it is not intuitive to explore any pairs of interactions among these variables. After fitting the model, a summary of the coefficients, test statistics, and Akaike Information Criterion (AIC) values are obtained in the summary of the fitted model. The model summary is attached in the Appendix.

The prediction results of the testing set are generated using the fitted model to evaluate the accuracy. A confusion matrix is also generated to show the number of true predicted and false predicted values. The matrix and the table of accuracy, as well as its interval, are shown below:

```
##        Criteria         Value
## 1      Accuracy  0.9345615045
## 2         Kappa -0.0007636877
## 3 AccuracyLower  0.9275043340
## 4 AccuracyUpper  0.9411270003
## 5  AccuracyNull  0.9996161965
```

```
##   Prediction True_Value_0 True_Value_1
## 1          0            0          339
## 2          1            2         4870
```

The final accuracy of the model is 0.9345615, and the AIC is almost 7000.

Conclusion

The summary of the model fitted is attached in the appendix.

Examining the summary of the model fitted, we can see that not all of the variables collected have explanatory value. LocalAuthorityCode does not seem to have significant explanatory power to the pass rate of food provider health checks. This is not surprising. Originally, we thought that due to varying socioeconomic circumstances among the boroughs of London, we could see a geographical contribution to health check pass rate where poorer neighborhoods were less likely to pass. However, the variable LocalAuthorityCode only encodes the arbitrary number designation of the borough and not its geographical location or socioeconomic status. If we wanted to explore this aspect further, more descriptive data such as household income of area or crime statistics will be much more useful to the model. As it currently stands this particular variable does not offer anything of interest.

Among the business types, Hospital/Childcare/Caring Premises, Other catering premises, and School/College/University were the categories of business types found significantly impacting the pass rate. Schools and Hospitals were found significant at level 0.01, while Other was found significant at level 0.05. All three of these categories had a positive effect on the logit transformed pass rate orders of magnitude larger than other categories. This result confirms our priors that public institutions are going to have the best compliance to public health standards. Hospitals and schools also interact with immunocompromised groups such as children and the elderly, so risks of poor hygiene have an outsized negative risk to the public. In terms of the Other category, the model serves as motivation to gather more data among these business types to offer more precise explanation.

Finally, Year and Month were found to both have significant negative effect on the logit transformed pass rate. This suggests that, over the course of a year and over the years, pass rate has declined. Originally we thought that the month component in decline could be that food hygiene standards worsen as London enters the holiday seasons, but given that the estimated change for the year (-4.883e-01) is roughly 12x that of the estimated negative change for the month (-4.648e-02), it seems that the change is the result of a continuous decline of pass rate, rather than some within-year seasonality. There could be many explanations to the decline of pass rate, one worth noting is that the range of data covers time both before and during the COVID-19 pandemic, when the food service industry experienced drastic changes such as additional pandemic safety restriction and standards. These unprecedented circumstances may be a contributing factor to the decline in pass rate.

Overall, we are satisfied with this model as it offers good explanations to the variables we are interested in and has high predictive power (0.935) on a large dataset. However, as mentioned above, more work is needed to draw conclusions regarding the geography of pass rate, the Other business category, and the effects of passing time and the COVID-19 pandemic.

Reference

- DATOTA. (2022). *UK Food Hygiene Rating Data (London)*[Data Set]. https://www.kaggle.com/datasets/datota/uk-food-hygiene-rating-data-london

Appendix - R Script

```r
# R code for the final project report
library(tidyverse)
library(ggplot2)
library(GGally)
library(rsample)
library(InformationValue)
library(caret)
# Read the dataset
data <- read_csv("./data/food_hygiene_rating_data.csv", col_names=TRUE)

# Clean the dataset
data_eda <- data %>%
  select(LocalAuthorityCode, RatingDate, BusinessType, RatingValue) %>%
  filter(RatingValue==0 | RatingValue==1 | RatingValue==2 |
           RatingValue==3 | RatingValue==4 | RatingValue==5) %>%
  mutate(Pass = ifelse((RatingValue >= 3), 1, 0)) %>%
  mutate(Year = format(RatingDate, "%Y") %>% as.numeric(),
         Month = format(RatingDate, "%m") %>% as.numeric(),
         Day = format(RatingDate, "%d") %>% as.numeric(),
         BusinessType = BusinessType <- as.factor(BusinessType)) %>%
  select(-RatingDate) %>%
  drop_na()
```

```r
# Visualization of the dataset
png("results/LAC_hist.png",width=800, height=600)
LAC_hist <- hist(data_eda$LocalAuthorityCode,
                 xlab="Local Authority Code",
                 main="Histogram of Local Authority Code")
dev.off()

options(repr.plot.width = 15, repr.plot.height = 8)
ggplot(data=data_eda, aes(x=factor(BusinessType))) +
  geom_bar(stat="count") +
  xlab("Count") +
  ylab("Business Type")+
  ggtitle("Bar Plot of Business Type")+
  theme(plot.title = element_text(hjust = 0.5))+
  coord_flip()
ggsave("./results/BT_bar.png")

ggplot(data=data_eda, aes(x=factor(Pass)), height=3) +
  geom_bar(stat="count") +
  xlab("Count") +
  ylab("Rating Value")+
  ggtitle("Bar Plot of Passing")+
  theme(plot.title = element_text(hjust = 0.5))+
  coord_flip()
ggsave("./results/RV_bar.png")
```

```r
# Split the dataset into training and testing sets
data_split <- initial_split(data_eda, prop = 0.75, strata = RatingValue)
training_data <- training(data_split)
```

```r
testing_data <- testing(data_split)
test_X <- testing_data %>%
    select(-Pass)
true_vals <- testing_data$Pass

# Fitting the models
mod <- glm(Pass~LocalAuthorityCode+BusinessType+Year+Month+Day,
           data=training_data, family="binomial")

# Model evaluation
mod_sum <- summary(mod)
pred <- predict(mod, newdata=test_X, type = "response")
cut_off <- optimalCutoff(true_vals, pred)
final_predict <- ifelse(pred > cut_off, 1, 0)
final_predict <- as.factor(final_predict)
true_vals <- as.factor(true_vals)
conf <- confusionMatrix(true_vals, final_predict)
conf_mat <- as.table(conf)
acc <- as.matrix(as.matrix(conf,what="overall")[c(1,2,3,4,5),])
write.csv(acc,file="./results/accuracy.csv")
write.csv(conf_mat, file="./results/conf_mat.csv")
```

Appendix - Model Summary

```
##
## Call:
## glm(formula = Pass ~ LocalAuthorityCode + BusinessType + Year +
##     Month + Day, family = "binomial", data = training_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7770   0.1972   0.3083   0.4462   0.7507
##
## Coefficients:
##                                                  Estimate Std. Error z value
## (Intercept)                                     9.549e+02  5.996e+01  15.925
## LocalAuthorityCode                             -5.367e-03  3.409e-03  -1.575
## BusinessTypeFarmers/growers                     1.101e+01  1.929e+02   0.057
## BusinessTypeHospitals/Childcare/Caring Premises 2.027e+00  5.934e-01   3.417
## BusinessTypeHotel/bed & breakfast/guest house   8.438e-01  6.410e-01   1.316
## BusinessTypeImporters/Exporters                -5.923e-01  8.271e-01  -0.716
## BusinessTypeManufacturers/packers               3.491e-01  5.982e-01   0.584
## BusinessTypeMobile caterer                      1.267e+00  6.280e-01   2.018
## BusinessTypeOther catering premises             1.635e+00  5.571e-01   2.935
## BusinessTypePub/bar/nightclub                   3.567e-01  5.527e-01   0.645
## BusinessTypeRestaurant/Cafe/Canteen             2.685e-01  5.312e-01   0.505
## BusinessTypeRetailers - other                  -1.300e-01  5.317e-01  -0.245
## BusinessTypeRetailers - supermarkets/hypermarkets 8.713e-01  6.003e-01   1.451
## BusinessTypeSchool/college/university           2.459e+00  6.373e-01   3.858
## BusinessTypeTakeaway/sandwich shop              2.016e-01  5.345e-01   0.377
## Year                                           -4.701e-01  2.974e-02 -15.805
## Month                                          -4.316e-02  9.066e-03  -4.760
## Day                                            -3.559e-04  3.844e-03  -0.093
```

```
##                                                Pr(>|z|)
## (Intercept)                                     < 2e-16 ***
## LocalAuthorityCode                             0.115351
## BusinessTypeFarmers/growers                    0.954493
## BusinessTypeHospitals/Childcare/Caring Premises 0.000634 ***
## BusinessTypeHotel/bed & breakfast/guest house  0.188023
## BusinessTypeImporters/Exporters                0.473951
## BusinessTypeManufacturers/packers              0.559481
## BusinessTypeMobile caterer                     0.043582 *
## BusinessTypeOther catering premises            0.003338 **
## BusinessTypePub/bar/nightclub                  0.518660
## BusinessTypeRestaurant/Cafe/Canteen            0.613304
## BusinessTypeRetailers - other                  0.806777
## BusinessTypeRetailers - supermarkets/hypermarkets 0.146674
## BusinessTypeSchool/college/university          0.000114 ***
## BusinessTypeTakeaway/sandwich shop             0.705990
## Year                                            < 2e-16 ***
## Month                                          1.93e-06 ***
## Day                                            0.926229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7608.9  on 15629  degrees of freedom
## Residual deviance: 6995.6  on 15612  degrees of freedom
## AIC: 7031.6
##
## Number of Fisher Scoring iterations: 12
```