JFK Data Analysis - Phase 2

Analysis Overview

This R Markdown file summarizes Phase 2 of the JFK complete data analysis. It presents the core R code used to rank each variable's influence on taxi_out. Most data preparation was done in SQL, and all effect-size calculations were performed in R using Pearson correlations and ANOVA for numeric and categorical predictors.

Setup

The following libraries were used:

```
library(dplyr)
library(lsr)
library(lubridate)

airport_data<-read.csv2("airport_data_SQL.csv",
                        header = TRUE, sep = ",")
weather_data<-read.csv2("weather_data_SQL.csv",
                        header = TRUE, sep = ",")
```

Data transformation

Three columns were created from timestamp to capture day, time-of-day, and hour.

```
airport_data<- airport_data %>%
        mutate(hour=hour(timestamp),
               time= case_when(
                       hour >= 5  & hour < 11 ~ "morning",
                       hour >= 11 & hour < 15 ~ "midday",
                       hour >= 15 & hour < 18 ~ "afternoon",
                       hour >= 18 & hour < 22 ~ "evening",
                       TRUE ~ "night"
               ),
               day= wday(timestamp,label =TRUE)

        )
weather_data$pressure <- as.numeric(weather_data$pressure)
```

Elimination

Column "condition" was removed due to inconsistent labeling. For example temperature for the "Fair" condition ranges from 19 to 61, as shown below:

```
sort(unique(weather_data$temperature[weather_data$condition=="Fair"]))
```

```
##  [1] 19 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
## [26] 46 47 48 49 50 51 52 53 54 55 57 59 61
```

Methodology

Spearman and Kendall were checked informally and showed no major differences, so the main analysis uses Pearson for numeric variables and ANOVA for categorical variables.

```
#
get_df<-function(x){if (x %in% names(airport_data)) {
        airport_data
} else {
        weather_data
}}
num_vars<-c("dep_delay","distance","departures","arrivals",
            "hour","temperature","dew_point","humidity",
            "wind_speed","wind_gust","pressure")
char_vars<-c("carrier","flight_code","destination",
             "time","day","wind")
num_results<- sapply(num_vars,function(x){
        df <- get_df(x)
        return(cor(df$taxi_out,df[[x]], method = "pearson"))
})
char_results<-sapply(char_vars,function(x){
        df <- get_df(x)
        f<-as.formula(paste("taxi_out ~ ",x))
        model<-aov(f, data=df)
        return(etaSquared(model)[1])
})
num_results<-data.frame(
        "variable"= names(num_results),
        "r2"= round(num_results^2*100,2),
        row.names = NULL)
char_results<-data.frame(
        "variable" = names(char_results),
        "r2" = round(char_results*100,2),
        row.names = NULL
)
results<- rbind(char_results,num_results) %>%
        arrange(desc(r2)) %>%
        mutate(r2= paste(r2,"%"))
head(results,10)
```

```
##       variable      r2
## 1  flight_code 11.14 %
## 2   departures  3.62 %
## 3      carrier  3.46 %
## 4  destination  2.98 %
## 5         wind  1.23 %
## 6         time  1.13 %
## 7    wind_gust  0.92 %
## 8          day  0.58 %
## 9  temperature  0.46 %
## 10    arrivals  0.42 %
```

Shared Coverage Check

A Type II ANOVA was used to see how much signal the top variables share with each other.

```
model<-lm(taxi_out~destination+carrier+flight_code+departures, data = airport_data)
overlap<-etaSquared(model, type = 2)
overlap <- data.frame( variable = row.names(overlap),
                       r2 = round(overlap[,1] * 100, 2) ,
                       row.names = NULL ) %>%
       arrange(desc(r2)) %>%
       mutate(r2 = paste0(r2, "%"))
overlap
```

```
##      variable    r2
## 1 flight_code 6.98%
## 2  departures 2.51%
## 3 destination 1.73%
## 4     carrier    0%
```

Notes

A full explanation of this phase is available in the Phase 2 documentation.