Week 10

Jon Cannaday

2024-02-27

Movie Theater Attendance

Introduction:

This paper serves as a comprehensive guide through the intricate process and challenges involved in crafting a Data Science model. It navigates through the evolution of a problem, rummages into data engineering methods, and comes to a climax in the generation of model predictions. Laced throughout the account are reflections and considerations, addressing insights gained and potential modifications.

The integration of Data Science methodologies can prove invaluable for theaters, offering insights for budgeting hours, conducting market research, and refining pricing strategies. Traditionally reliant on managerial intuition, data science opens avenues for informed decision-making based on data and mathematical models.

Problem Statement:

Originating from the need to forecast attendance for upcoming movies, the model's reach goes beyond attendance predictions. The emphasis on predicting differing outcomes emphasizes its utility in assisting strategic decisions. Within the movie industry, this predictive capacity enhances the precision of budgeting employee hours.

Problem Addressed:

The problem is addressed by firstly, unpacking the underlying reason held by each record within a table. These were identified by HASH values made up of theater, movie ID, and print ID for the movie dimension. Transaction-level fact tables provide the foundation for attendance derivation. Addressing challenges, such as incorrect date entries, involves creative solutions, like leveraging showtime tables to construct makeshift movie tables with accurate release dates.

Highlighting the objectivity of data is pivotal, stressing the alignment with reality. Each record mirrors a real-life interaction, and the data aims to objectively represent these interactions, forming the bedrock for decision-making.

Aggregating data to a consistent level of detail posed challenges, met with the utilization of SQL Common Table Expressions (CTEs). These expressions group each table to a specific detail level, close to finding the greatest common denominator in mathematics. Opting for the highest data view facilitates grouping by attributes such as location, date, movie, and customer.

The data flow is designed for minimal change effort, with provisions for easy column addition and dynamic column categorization in Python. Variables and a for loop ensure a cohesive approach, allowing the model to seamlessly adapt to changing experimental variables.

Analysis:

The exploration starts with Exploratory Data Analysis (EDA), searching each table for insights into records and data generation. The linking of data using keys, exploration of frequency tables, histograms, and descriptive statistics, and leveraging scatter plots and correlations aid in feature selection. Slicing and dicing the data involved changing aggregation levels, experimenting, and applying models at varying granularity, such as theaters, dates, and movie genres.

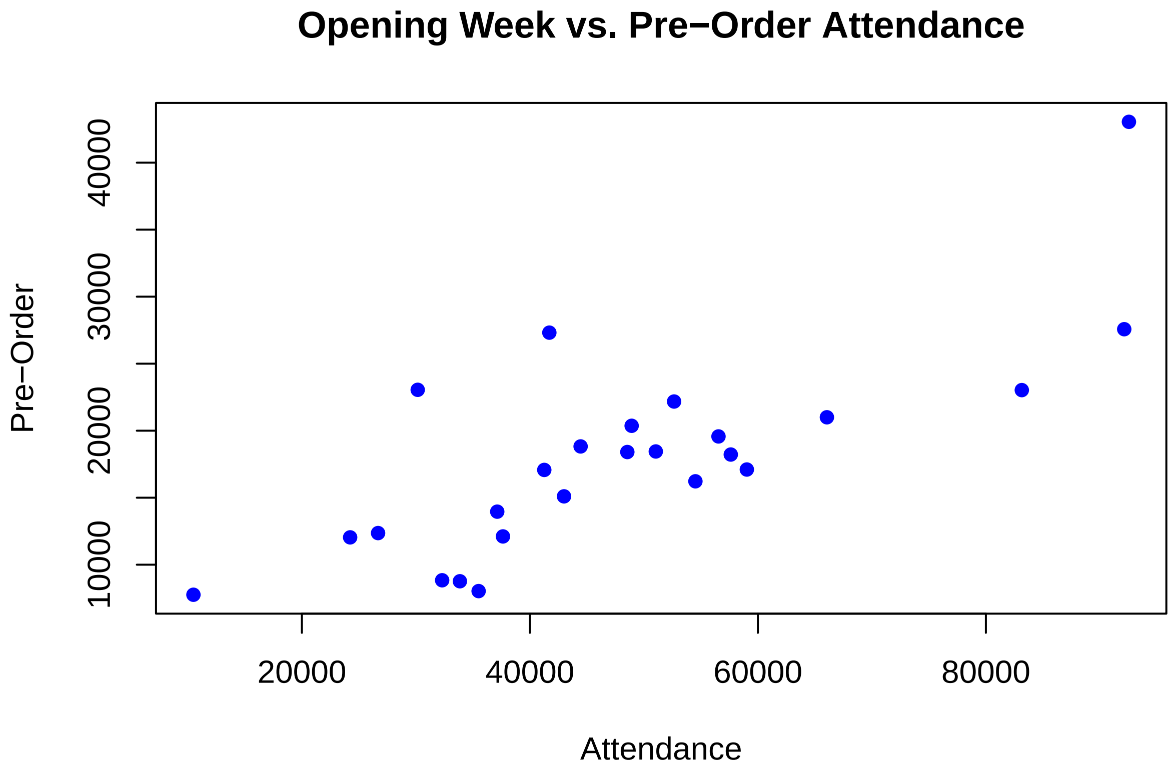
```
library(readxl)

# Data
path <- "C:/Users/Owner/OneDrive/Desktop/DataScience/2023_Winter_Statistics/"
file <- 'Movie_Theater_Data.xlsx'
filepath <- paste(path, file, sep = "")
data <- read_excel(filepath)
```

```
## New names:
## * ‘DateIsWinter‘ -> ‘DateIsWinter...20‘
## * ‘DateIsWinter‘ -> ‘DateIsWinter...21‘
```

```
head(data, 5)
```

```
## # A tibble: 5 x 60
##   TheaterNumber MonthDate           OpenAttendance PreOrderAttendance
##           <dbl> <dttm>                       <dbl>              <dbl>
## 1           231 2023-01-01 00:00:00          24233              12041
## 2           231 2021-12-01 00:00:00          44450              18826
## 3           231 2022-02-01 00:00:00          41262              17072
## 4           231 2022-01-01 00:00:00          26682              12360
## 5           231 2021-02-01 00:00:00          35503               8031
## # i 56 more variables: DateIsJanuary <dbl>, DateIsFebuary <dbl>,
## #   DateIsMarch <dbl>, DateIsApril <dbl>, DateIsMay <dbl>, DateIsJune <dbl>,
## #   DateIsJuly <dbl>, DateIsAugust <dbl>, DateIsSeptember <dbl>,
## #   DateIsOctober <dbl>, DateIsNovember <dbl>, DateIsDecember <dbl>,
## #   DateIsSpring <dbl>, DateIsSummer <dbl>, DateIsFall <dbl>,
## #   DateIsWinter...20 <dbl>, DateIsWinter...21 <dbl>, DateIsNearNewYears <dbl>,
## #   DateIsNearValentines <dbl>, DateIsNearEaster <dbl>, ...
```

```
plot(data$OpenAttendance, data$PreOrderAttendance, main = "Opening Week vs. Pre-Order Attendance",
     xlab = "Attendance", ylab = "Pre-Order", pch = 16, col = "blue")
```



```r
cor_matrix <- cor(data[, c("OpenAttendance", "PreOrderAttendance")])
print(cor_matrix)
```

```
##                    OpenAttendance PreOrderAttendance
## OpenAttendance          1.0000000          0.7679784
## PreOrderAttendance      0.7679784          1.0000000
```

Implications:

Identifying a high correlation between presales and opening weeks attendance per movie has operational implications for theater managers, influencing staffing decisions. A brief exploration of a 3D correlation matrix concludes that two-dimensional correlations suffice. The organization gains insights into data operations and their broader implications.

Limitations:

Limitations include the reliance on data from a single theater, warranting future iterations including the entire country. The unavailability of external transaction-level data and additional features like GPS API-based location specifics and astrological dates is acknowledged. Certain information, like the need for additional features using an API and implied logic in release dates, is not self-evident.

Conclusion:

In conclusion, the data exhibits linear trends, particularly in presales and opening weeks sales. The data engineering framework proves adaptable to diverse business angles and experimental trends, showcasing presales as reliable indicators of opening attendance. This information serves stakeholders for informed decision-making. Following initial regressions, the plan is to further dissect the data, experiment with different methods, and include more features to enhance the model. Continuous improvement and exploration of various analytical methods are central to a comprehensive analysis.