Queensland University of Technology

IFN701 PROJECT – SEMESTER I, 2019

PROJECT NAME:

EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).

STUDENT NAME:	Victor Manuel Villamil Romero
SUPERVISOR:	Dr. Dimitri Perrini
STUDENT ID:	n10080236
PROGRAM:	Master of Information Technology
MAJOR:	Data Science
PROJECT CATEGORY:	Development Project

Executive Summary

Travelling across several time zones might have a negative impact over the performance of professional team players and their teams (Youngstedt & O'Connor, 1999) (Mcguckin, Sinclair, Sealey, & Bowman, 2014). The evaluation and understanding of travelling across the time zone which may affect the optimal performance whether offensive or defensive of professional teams can pose a challenging task for researchers and people in sports industry. The objective of the present report is to outline the process of data analysis developed over the offensive statistics of baseball teams in games from 1975 to 2018 from the Major League Baseball (MLB) of United States, where the variable of jetlag was included.

The present project is an extension of previous efforts done in the area of jet lag analysis over player performance (Song, Severini, & Allada, 2017). However, the scope of the present report comprises the usage of two different linear models where the effect of jet lag was considered as both boolean and categorical. It also will focus on 6 offensive measurements such as Batting Average (BA), On Base Performance (OBP), Slugging Percentage (SLG), Offensive Average (OA), Runs scored, and Offensive performance Average (OPA). Moreover, the value of jetlag was calculated by considering the number of time zones crossed where a compensation factor was added to consider the adaptation capacity of human body. In addition, the ELO rating to represent the relative skill level of each team was considered as part of the general scope. The method followed by the project is compound by the followed by Song , Severini and Allada (2017) and the Knowledge Discovery Process in Databases (Liu, Tian, & Zhang, 2010), where the stages of data Selection, Pre-Processing, Transformation, Data Analysis and Interpretation/Evaluation were implemented and reflected. Finally, four main outcomes were produced, a python code containing the initial data transformation, a R code containing the statistical data analysis executed over the dataset, a presentation of the results and the analysis of the results as a report format.

As a result of the analysis, was possible to conclude that jet lag effects do affect offensive performance whether by considering it as a boolean representation or a categorical one. However, the results varied across models where different importance among direction (eastward or Westward) and home or away team position were observed.

REVISION HISTORY

The following table will register the modifications developed in the present project plan with the aim to keep historical register.

Version	Format	Date Created	Description	Modification Performed
0.0	Doc	01/05/2019	Creation of Document	Creation
0.1	Doc	10/05/2019	Inclusion of additional sections	Edition
0.2	Doc	11/05/2019	Inclusion of additional sections	Edition
0.3	Doc	15/05/2019	Inclusion of additional sections	Edition
0.4	Doc	20/05/2019	Inclusion of additional sections	Edition
0.5	Doc	22/05/2019	Inclusion of results and discussion	Edition
0.6	Doc	24/05/2019	Release of first version	Draft Creation
0.7	Doc	30/05/2019	Finishing final version	Finishing
1.0	Doc	02/06/2019	Edition and final amendments	Releasing.

Table of Content

LIST OF FIGURES

Introduction

Background and problem statement

Travelling across several timezones might play a significant role in professional team player performance. While several investigations regarding performance analysis have been focused on a time constraint variables such as recuperation time, leisure time and competition time (Richard, R, 2018), there is limited research on the impact of travels across different timezones during a regular season in professional sports. Previous studies in the area of sport and exercise physiology have defined jetlag as the group of consequences produced by the disruption of circadian rhythm when several time zones are crossed occurring after long-haul flights (Birch, K., MacLaren, D., & George, K, 2005). Those symptoms may include among others lack of alertness, impaired motor performance, gastrointestinal disturbance and fatigue which might affect the normal behaviour of players.

Youngstedt, O'Connor (1999) and Mcguckin, Sinclair, Sealey, and Bowman (2014), support that teams and players might be negatively affected after transmeridian flights, reducing their performance and affecting their results. Moreover, other studies claim that there are significant differences between the jetlag effects of travelling westward and eastward (Monk, T., Buysse, D., Carrier, J., & Kupfer, D. (2000) which may also affect significantly the results of professional players. However, the selection of an adequate scenario to measure the jetlag effects on performance based on a data analytics approach is not an easy task. This analysis is especially difficult due to the importance of meaningful statistics, a high demanding physical effort and a short time frame between games in different time zones. According to (Baumer & Zimbalist., 2013), baseball can be considered as a sport with a wide variety of meaningful statistics, where several previous studies (Bennett & Flueck, 1983) (Pankin, 1978) have introduced measurements to devel the more relevant relationships between the events during the games. It also accounts not only for a physically demanding nature but also for a short timeframe between games in many leagues across the world.

The present project takes advantage of freely available data from 1871 to 2018 of baseball games from the Major League Baseball (MLB) in the United States collected by the Retrosheet organization https://www.retrosheet.org/gamelogs/index.html. This data contains information about each one of the games played by each team during each season including all the events occurred during the match. This data meets the statistical input requirements to generate the offensive and defensive measurements required. It also contains general information regarding time, location and conditions for each game which allow crossing reference information about travels across different time zones including the direction to infer the jetlag variable to be used. However, in order to add the variable of jetlag, only data from 1975 and onwards were included in the present analysis.

Aims and Objectives

The aim of the present project is to evaluate the significance between the presence of jetlag effects measured as a function of the number of timezones crossed and a defined group of offensive performance of teams from Major league Baseball (MLB) which will be extended in previous sections.

In order to support the aim previously defined, the project set the following objectives to be achieved:

1. Following the indication of previous studies (Song, Severini, & Allada, 2017), for 5 weeks of the project execution, select, transform and extract the data of baseball games from 1974 to 2018 provided by Retrosheet organization (Retrosheet, 2018) allowing the calculation of offensive statistics including Runs scored, Batting average, On-base % and Slugging %, the jetlag effect as a direction and value depending of number of time zones crossed for each baseball team in subsequent games, the differences among the relative skills of each team measured as the ELO value.

2. For 3 weeks of the project, transform the final dataset accounting the variables previously defined and apply statistical analysis techniques such as multivariate linear regression analysis in order to devel the significance among the variables representing the presence of jet lag and the offensive statistics defined previously.

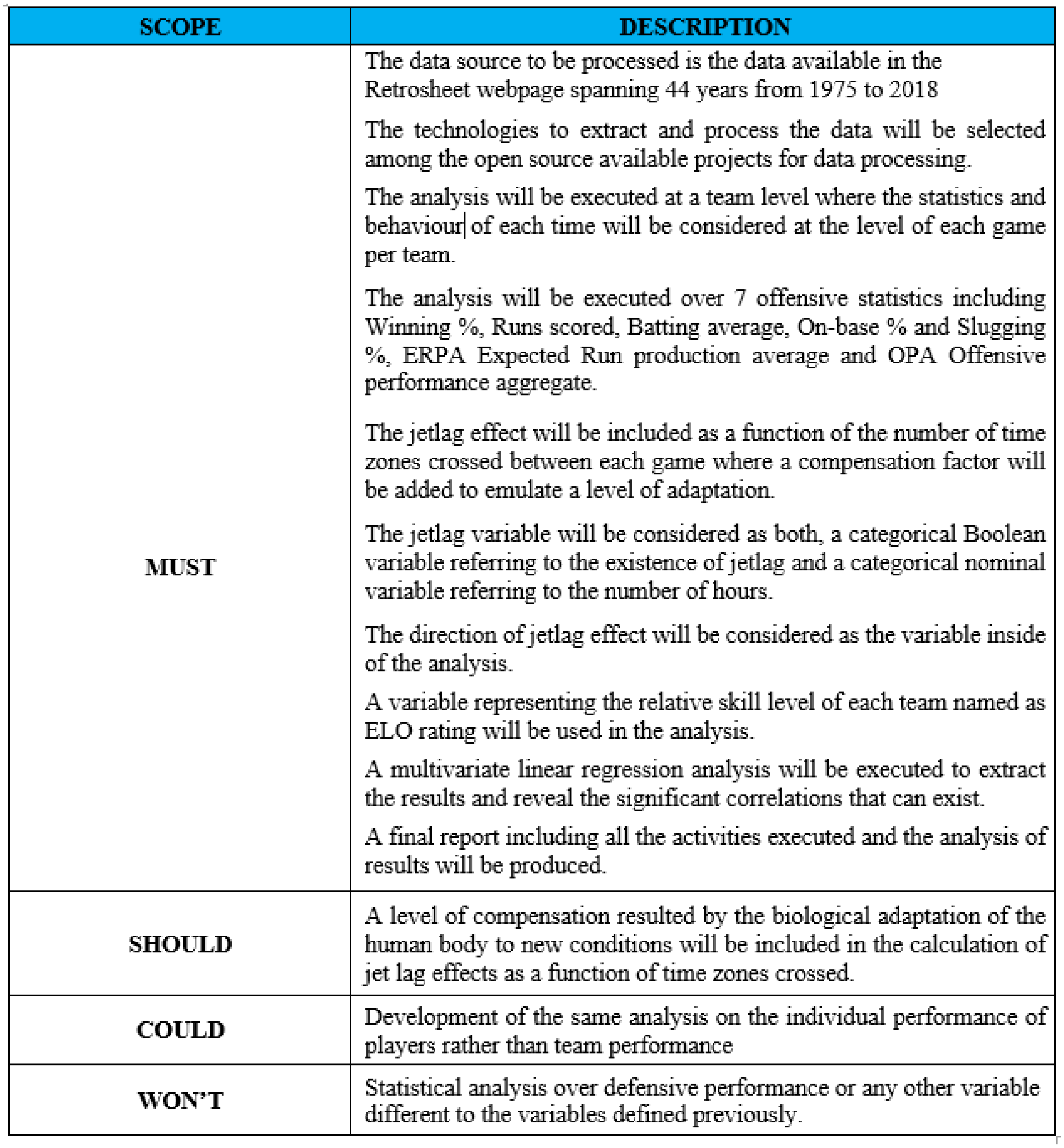
3. For 1 week of the project, organise and produce three main deliverables including the developed prototype, a presentation of the results and a project report showing the results gathered from the analysis.

Brief overview of Methods

During the execution of the project, three main stages were covered. During the first stage, an investigation about the context of the project was executed by gathering information regarding baseball performance analysis, baseball offensive statistics, jetlag analysis (including compensation and measurement), the relative skill of baseball teams and previous studies. During this stage, the different considerations, modifications, compensations and approaches to be applied over the data were analyzed and defined. Then, during the implementation stage, the Knowledge Discovery from Databases (KDD) process was followed to approach the data analysis project. Each one of the stages conforming this process will be extended in further sections. Finally, a results presentation and evaluation stage were applied to generate and compile the final outcomes which are also extended further sections.

Recap of Scope

The scope of the present project was defined by following an agile prioritization technique known as MoSCoW, which allowed to define three different levels of priority. Each level and the elements included for each one are shown in the following table



SCOPE	DESCRIPTION
MUST	The data source to be processed is the data available in the Retrosheet webpage spanning 44 years from 1975 to 2018 The technologies to extract and process the data will be selected among the open source available projects for data processing. The analysis will be executed at a team level where the statistics and behaviour of each time will be considered at the level of each game per team. The analysis will be executed over 7 offensive statistics including Winning %, Runs scored, Batting average, On-base % and Slugging %, ERPA Expected Run production average and OPA Offensive performance aggregate. The jetlag effect will be included as a function of the number of time zones crossed between each game where a compensation factor will be added to emulate a level of adaptation. The jetlag variable will be considered as both, a categorical Boolean variable referring to the existence of jetlag and a categorical nominal variable referring to the number of hours. The direction of jetlag effect will be considered as the variable inside of the analysis. A variable representing the relative skill level of each team named as ELO rating will be used in the analysis. A multivariate linear regression analysis will be executed to extract the results and reveal the significant correlations that can exist. A final report including all the activities executed and the analysis of results will be produced.
SHOULD	A level of compensation resulted by the biological adaptation of the human body to new conditions will be included in the calculation of jet lag effects as a function of time zones crossed.
COULD	Development of the same analysis on the individual performance of players rather than team performance
WON'T	Statistical analysis over defensive performance or any other variable different to the variables defined previously.

Figure 1. Description of final Scope.

Key deliverables

The key outputs of the projects are going to be summarized in three main digital deliverables:

1. **Development**: The source code for the data preparation in python and the data analysis in R.
    a. **Python Code:** It contains all the functions and classes used during the extraction, preprocessing and transformation stages.
    b. **R code:** It is used to execute the statistical data analysis and the visualization of results.

2. **Results Report:** A document containing all the information related to the project including an understanding of previous literature, project methodology, results and conclusions.
3. **Presentation:** A file containing the final presentation of the project and also the principal and most representative results. The format of this presentation can be provided in .pdf or .ppt.

The significance of the present project can be reflected not only in the practice area but also in the researching area. In the practice, team owners, coaches and professionals working in the market of sports can benefit from the findings by increasing their awareness related to the effects of travelling across time zones over the professional players allowing them to generate strategies to reduce the impact and avoid any negative effect over the team results. On the other hand, researches in the area of neuroscience, sleep and human performance can take advantages of the results to compare, judge and validate their analysis related to this area.

1. Environmental Scan

Previous efforts have been done in the area of jet lag impact over professional players performance. Spanning from general medical analysis as the published by Youngstedt and O'Connor (1999) to discipline-focused as the research of European Journal of Sport Science developed over rugby players (Mcguckin, Sinclair, Sealey, & Bowman, 2014). However, the results gathered from those analyses have not been completely clear to conclude a direct relationship between air travels and professional performance. Both pieces of research have suggested that more rigorous investigations should be done to establish whether or not crossing several time zones might impair athletic performance. Therefore, further efforts surrounding the experiments followed by those researches and other should be developed, where different variables, assumptions and particularities surrounding the approaches used by these studies might be modified to increase the awareness in this area.

First of all, the concept of circadian rhythm used to evaluate the jet lag impact should be introduced. It is defined as a natural process to regulate psychological functions which oscillate from peak to lowest throughout a period of 24 hours (Smith, Guilleminault, & Efron, 1997). Some main examples of these circadian rhythms are the sleep-wake cycle, cognitive and physical performance. Based on these rhythms, some studies argue that there are significant fluctuations in human performance related to the time of the day (Klein, Wegmann, & Athanassenas, 1976). They suggest that some athletes experiment a peak on athletic human performance during the afternoon while other experience improvements during the night. Other findings also evidence that alterations or disruptions of circadian rhythm as the produced by travelling across time zones may produce adverse effects on performance and psychological responses (Graeber, Kryger, Roth, & Dement, 1994) (Redfern, Minors, & Waterhouse, 1994).

By supporting the findings related to affections over circadian rhythms, previous research has evaluated how these variations produced by crossing several time zones may be significant regarding the direction of travelling or the home-away advantage of the team. Some studies have focused on the differences between the likely impact of jetlag effect given the direction of travels either east or west, which can be understood as the forward or backward modification of internal clock effects (Fowler, et al., 2017). Results indicate that travelling eastward evidenced more significant effects over physical performance reflected on a detrimental effect on fatigue, sleep and motivation. It also concluded that these effects were shown within 72 hours after the travels. On the other hand, studies over baseball games result considering the effect of home or away after crossing several time zones shown that home position of teams was more significant than away teams (Recht, Lew, & Schwartz, 1995).

Some special attention should be placed over the previous results of Fowler et al (2017) regarding the timeframe within the effects showed significantly. It implies that the effects should be considered as temporal dependant where the relative effects vary in relation to the number of days. Some studies have defined a 1-hour reduction per every 24 hours of timeframe after the alteration (Wever, 1966) (Takahashi, et al., 1999). This approach may set a reference point where the ability of auto compensation of human bodies is taken into consideration. Moreover, this approach allows filtering the number of instances where a real presence of jet lag may appear, focusing the analysis only in the really relevant situations.

As was mentioned before, previous research has suggested evaluating different variables that can be related to player performance. Following this suggestion, a variable to rate the relative skill level of each team might be considered. More specifically, the significant effects coming from the differences among the level of skills from teams known as Elo rating will be evaluated. The Elo

Rating system is broadly accepted in several sports and disciplines due to the advantage of considering not only the history of the team or player but also the factor time (Lehmann & Wohlrabe, 2017). It was originally used in chess players but it has spread to sports such as tennis, baseball and so on. According to Lehmann and Wohlrabe (2017), the calculation of Elo rating comprises two steps, the first step calculates the expected score while the second step updates the player's rating. Glickman and Jones (1999) describe the method used to calculate the Elo rating for players in chess games which can be extended to other disciplines by applying the same formulas.

Following the previous precedents, the present project will provide an additional evaluation of the relationship between circadian misalignment and performance. It will develop a statistical data analysis over information from real games influenced by travelling across several time zones and the correspondent team performance. The analysis developed during the present project will be aligned to previous research developed by Song, Severini and Allada, R.(2017) over data from baseball sport. In this study, data from the major league baseball of the United States provided by the Retrosheet organization (Retrosheet, 2018) was used, where results gathered shown significant effects after eastward travels and home teams. Results aligned to the findings of Fowler, et al (2017) and Recht, Lew, and Schwartz (1995). Moreover, the researchers concluded that the results have shown focalized effects of travelling across time zones on athletic performance. This is particularly interesting given the previous results in the medical analysis (Youngstedt & O'Connor, 1999) and other disciplines (Mcguckin, Sinclair, Sealey, & Bowman, 2014). Therefore, an extension over this study may be executed where the same methodology can be used and more meaningful variables can be added.

Finally, one important consideration during previous efforts in the area of the baseball team and player performance has been the definition of meaningful statistics that represent accurately the likely variations. Some popular classifications of a statistic such as defensive and offensive ones are commonly used (Song, Severini, & Allada, 2017). However, several studies in the area of baseball performance as the developed by Pankin, M. D. (1978) and Bennett, J., & Flueck, J. (1983) suggest the inclusion of additional statistics. In this way, the offensive Average (OA) and Slugging percentage (SP) are recommended as estimators of performance which generally produce similar rankings (Bennett & Flueck, 1983). Moreover, the OPA (Pankin, 1978) which includes the contribution of stolen bases over the rise of expected runs done by the batter is also recommended as one of the offensive statistics to be evaluated.

2. Project Methodology

The methods followed during the project were based on a combination of previous research in the area of data analytics (Liu, Tian, & Zhang, 2010)and jetlag performance (Song, Severini, & Allada, 2017). They can be summarized in three general stages named as Analysis and understanding, Implementation and Results Evaluation.

2.1. Analysis and Understanding

The objective of the first stage is to set up the project before the implementation stage starts. It covers four sections including the design of the program architecture to be developed, the approach to be used during the jet lag assessment, the baseball offensive statistics to be considered under the analysis and the new variable regarding the relative skill set of teams.

2.1.1. Programming Approach

Five principal components were defined. The objective of each one of the components and additional information is presented below:

- **Main Program:** This component will contain the general logic of the program from where all the other components will be called.

- **Data Extractor:** This component will contain the methods required to execute the web scrapping into the Retrosheet webpage (Retrosheet, 2018) and to transform the data extracted into structured tables.

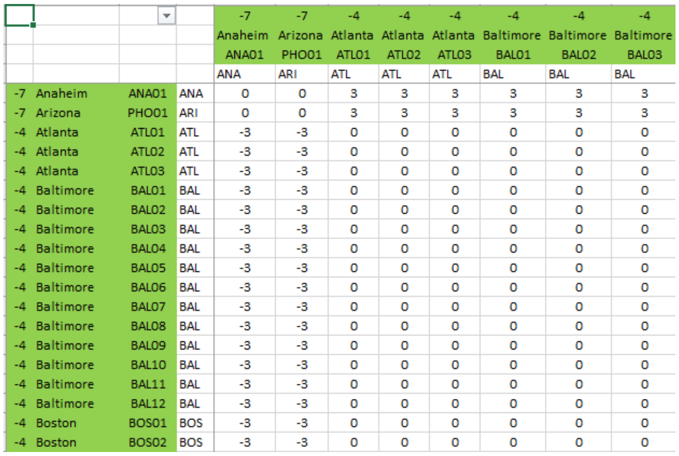
- **Data Importer:** This component will contain the methods required to import the data created to handle the external variables as the crossed time zones, Elo performance and Cfip. It will be developed in python.

- **Data transformer:** This component will contain all the methods to be applied over the data extracted in order to create the final dataset required for the analysis. will produce as output, one unique file named *df_Final_Dataset_1975_to_2018*. It will be developed in python.

- **Data Analyzer :** This component will contain all the statistical data analysis to be executed over the final dataset. It will allow to analyze visually and statistically the data provided by the python program. It will be created in R.

2.1.2. Jet Lag Assessment

First of all, the different combinations of time zones (TZ) crossed across games locations needs to be mapped. To do that, the project created a matrix where each stadium, the city where the stadium is located, and the time zone of the city were cross-referenced with each other. This file is provided in the zip folder with the code as *timesZonesDayLight_Full.xls* and the image below shows its overall structure:



				-7	-7	-4	-4	-4	-4	-4	-4
				Anaheim	Arizona	Atlanta	Atlanta	Atlanta	Baltimore	Baltimore	Baltimore
				ANA01	PHO01	ATL01	ATL02	ATL03	BAL01	BAL02	BAL03
				ANA	ARI	ATL	ATL	ATL	BAL	BAL	BAL
-7	Anaheim	ANA01	ANA	0	0	3	3	3	3	3	3
-7	Arizona	PHO01	ARI	0	0	3	3	3	3	3	3
-4	Atlanta	ATL01	ATL	-3	-3	0	0	0	0	0	0
-4	Atlanta	ATL02	ATL	-3	-3	0	0	0	0	0	0
-4	Atlanta	ATL03	ATL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL01	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL02	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL03	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL04	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL05	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL06	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL07	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL08	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL09	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL10	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL11	BAL	-3	-3	0	0	0	0	0	0
-4	Baltimore	BAL12	BAL	-3	-3	0	0	0	0	0	0
-4	Boston	BOS01	BOS	-3	-3	0	0	0	0	0	0
-4	Boston	BOS02	BOS	-3	-3	0	0	0	0	0	0

Figure 2. Time Zones crossed matrix.

Then the information regarding the number of time zones crossed and the stadium is extracted in a file named *timesZonesDayLight_site.xls* which is the one fed into the code. At this point, the Jet lag equation can be defined as follows:

$$Jetlag = TZ \qquad (1)$$

After gathering the number of time zones crossed, a compensation (C) value of 1 hour per every 24 hours of staying in the new time zone is applied (Wever, 1966) (Takahashi, et al., 1999). For this project, the assumption that the team travels from the current location to the new location immediately after finishing each game will be used to facilitate the analysis. Therefore:

$$C = days(Date_{Game} - Date_{Arrival}\ ) \qquad (2)$$

Subsequently, the jet lag equation can be updated:

$$Jetlag = (TZ - C) \qquad (3)$$

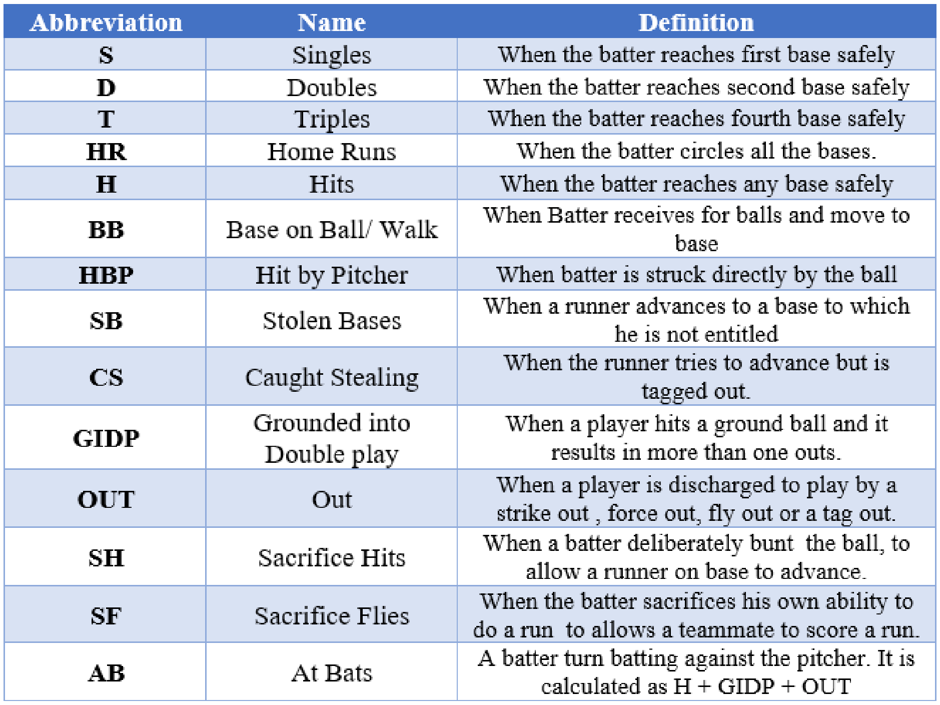
On the other hand, the effect of previous games should be considered into the actual equation. This effect can either reduce the overall number of time zones crossed if the team is coming back to the direction from where they travelled or increase the overall effect if the team is moving again towards the same direction. Therefore, the equation for jet lag measurement is defined as:

$$Jetlag_{Compensed} = R + (TZ - C) \qquad (4)$$

Where R represents the remaining jetlag from previous games, TZ represents the time zones crossed in the context of the actual game and C the compensation applied.

2.1.3. Offensive Statistics

During the present project, 6 offensive statistics will be selected to be used during the analysis. These six statistics were selected based on previous studies and other recommendations coming from the literature reviewed. In order to define the statistics selected, a nomenclature for the possible events occurring in a baseball game is shown below:



Abbreviation	Name	Definition
S	Singles	When the batter reaches first base safely
D	Doubles	When the batter reaches second base safely
T	Triples	When the batter reaches fourth base safely
HR	Home Runs	When the batter circles all the bases.
H	Hits	When the batter reaches any base safely
BB	Base on Ball/ Walk	When Batter receives for balls and move to base
HBP	Hit by Pitcher	When batter is struck directly by the ball
SB	Stolen Bases	When a runner advances to a base to which he is not entitled
CS	Caught Stealing	When the runner tries to advance but is tagged out.
GIDP	Grounded into Double play	When a player hits a ground ball and it results in more than one outs.
OUT	Out	When a player is discharged to play by a strike out , force out, fly out or a tag out.
SH	Sacrifice Hits	When a batter deliberately bunt the ball, to allow a runner on base to advance.
SF	Sacrifice Flies	When the batter sacrifices his own ability to do a run to allows a teammate to score a run.
AB	At Bats	A batter turn batting against the pitcher. It is calculated as H + GIDP + OUT

Figure 3. Abbreviation events baseball games.

- **Offensive Average (OA):** This statistic is defined as:

$$OA = \frac{S+2D+3T+4HR+W+SB}{AB+W} \qquad (5)$$

- **Offensive Performance Average (OPA) :** This statistic is defined as:

$$OPA = \frac{S+2D+2.5T+3.5HR+0.8(W+HBP)+0.5SB}{AB+W+HBP} \qquad (6)$$

- **Batting Average (BA) :** This statistic is defined as:

$$BA = \frac{H}{AB} \qquad (7)$$

- **On Base Performance (OBP) :** This statistic is defined as:

$$OBP = \frac{H+W+HBP}{AB+W+HBP+SF+SH} \qquad (8)$$

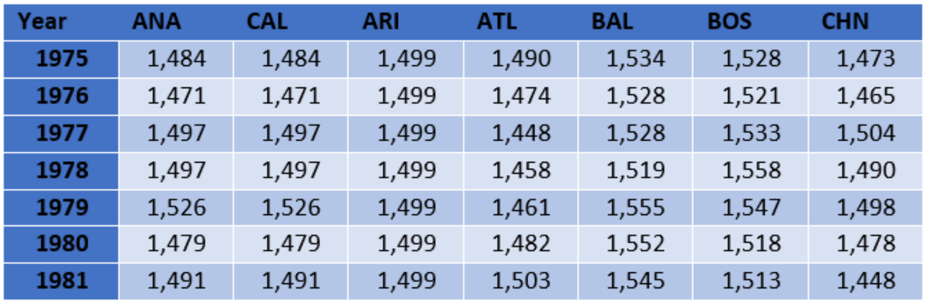
- **Slugging Percentage (SLG) :** This statistic is defined as:

$$SLG = \frac{S+2D+3T+4H}{AB} \qquad (9)$$

- **Runs Scored:** This statistic is defined as the number of complete passes through each base done during each game.

2.1.4. Relative skills (Elo).

In order to extend the scope of previous studies, a variable related to the relative skills for each team will be introduced. This value is calculated by team per year with data provided by the major league baseball. Therefore, the project created a matrix file to be fed to the program following the structure below:



Year	ANA	CAL	ARI	ATL	BAL	BOS	CHN
1975	1,484	1,484	1,499	1,490	1,534	1,528	1,473
1976	1,471	1,471	1,499	1,474	1,528	1,521	1,465
1977	1,497	1,497	1,499	1,448	1,528	1,533	1,504
1978	1,497	1,497	1,499	1,458	1,519	1,558	1,490
1979	1,526	1,526	1,499	1,461	1,555	1,547	1,498
1980	1,479	1,479	1,499	1,482	1,552	1,518	1,478
1981	1,491	1,491	1,499	1,503	1,545	1,513	1,448

Figure 4. Elo matrix created

It means that from the perspective of the principal team, the net difference between the Elo values of each team will be included for each game. The following equation is created to represent this inclusion:

$$NetElo = ELO_{TeamA} - ELO_{TeamB} \qquad (10)$$

Where Team A represents the team that is being used as principal team and Team B the against team. In this way, when the principal team accounts a higher relative skill level, the Net Elo will be positive whereas when the team against accounts for a higher value, the Net Elo will be negative.

2.2.Implementation

This stage will be aligned to the knowledge discovery database process defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

The image below shows the general KDD process:

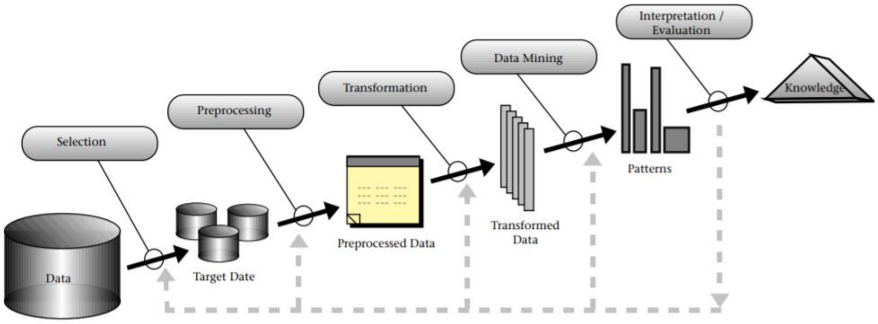


Figure 5. Knowledge Discovery Process in Databases (Liu, Tian, & Zhang, 2010)

2.2.1. Data Selection (Collection)

The activities on this stage will be found in one independent file named *Wrangling_Datasource.py*. The way as Retrosheet foundation made available the data and how it can be used for the project was understood. It consisted on:

- Downloadable .zip files per each year containing one file per team from both leagues (national and American). Each .zip folder is available through a public webpage. The project took advantage of the patterns found in the sequence of Urls created by Retrosheet to make the .zip folders available. The URL follows the structure of: https://www.retrosheet.org/events/XXXXeve.zip where XXXX represents the year under evaluation.

- Inside of each folder, there was a text file named TEAMXXX which contained a list of names for each team who played during the season. The project uses this information to harvest the files needed from the .zip due to each team accounts for a file where all the information from the season is contained.

- Each file is filled with rows representing in order the ID of the game, general information about the game, information about the roster of the team for the game and detailed information about each event during the game. These sequential structures were used in group with the descriptions provided by Retrosheet organization to arrange each category of information into a structured data type. In this way, 6 different tables were created per each year containing the correspondent information for all the teams during the season. Inside the program, six different dictionaries were created by using the years as keys and each table as values to this information.

2.2.2. Data Pre-Processing

The objective of this stage is to remove the noisy data and outliers. However, after extracted the data from Retrosheet webpage, the data was inspected and no missing or outlier values were found. On the other hand, in further stages some high outliers were included when the jetlag value was introduced. They were handled by imputation approach replacing them by the maximum value.

2.2.3. Data Transformation

The information contained inside the table play (one row per event in each game) can be considered as the most relevant, especially the information contained in the event field. The structure and meaning of each character on this field are described in Retrosheet webpage. The project used this information to extract each one of the events happened during each game which is aligned to the variables defined previously in the offensive statistics.

In the next transformation, the information of each game is grouped in two rows, grabbing only the information correspondent to each team. Each row of events happened in each game accounts for one column which identifies the team. This field is used to group all the events of each game in two rows, one from the perspective of the team A and the other from the perspective of team B. Each row will have the sum up of all the variables mentioned before from the perspective of each team per game. Finally, as these variables are related to the ones defined in the section of offensive statistics, the six offensive statistics previously defined are calculated and the columns with the variables of the event are dropped.

Then, the information about the site and date of game and site and date of the previous game played by each team is inserted. From these new variables, the number of time zones crossed (By using the matrix previously introduced) is calculated and the number of compensation days (using the assumption previously mentioned) are calculated. Then, per each team, the information regarding jetlag for the team against is included in the same row. Finally, the variable correspondent to Elo of each team and team against is included to each row, creating the final dataset

2.2.4. Data mining/Data Analysis

This stage will be developed in R where a file named *Data_Analyzer.R* will be created. During this stage, two principal analysis will be executed. The first analysis corresponds to visualize the data through meaningful plots which allow us to understand the nature of the dataset created. Some of the behaviors under evaluation are the balance between the categorical variables to be used in the analytical model and the normality of numerical variables. Moreover, the distribution of each offensive statistic across the levels of the jet lag variable will be evaluated for both the Boolean approach and the categorical approach.

The second technique to be executed will be a multivariate linear regression analysis. This technique will help to describe the relationship between the explanatory and each outcome variable. It also will follow the previous studies related to jet lag impact over offensive performance where the same technique was executed (Song, Severini, & Allada, 2017).

So, the model used will be:

$$y_i = \beta_1 X_1 + \beta_2 X_2 \ldots \ldots . + \beta_n X_n + \varepsilon_i \qquad (11)$$

Where β represents the coefficients accounted for each variable under evaluation and ε_i describes the leftover variability. During this project, a value of $\alpha = 0.05$ and 95% of confidence intervals will be used.

For this stage, two experiments will be defined where three variables including Venue (Home or Away), Direction of jet lag (Eastward or Westward) and Jet lag will be evaluated. Each experiment is extended below:

1. **Boolean Approach:** During this experiment, the variable jet lag will be evaluated as a Boolean data type. Two variables representing the possible combinations between venue and direction for each team will be inserted which will account the levels *HomeEast*, *HomeWest*, *AwayEast and AwayWest.* However, when hot encoding be applied over these variables, as both account for the same levels, only four columns (one per each level) should be created.

2. **Categorical Approach:** During this experiment, the variable jet lag will be evaluated as a categorical data type. The variables Venue and Direction will be assessed as in the Boolean approach. Moreover, a variable representing the level of final jet lag will be also included (-1,1,1,2). When hot encoding is applied, as the three variables account for the same levels from the perspective of each team, only 8 columns (one per each combination) should be considered (*HomeEast2*, *HomeEast1*, *HomeWest-2*, *HomeWest-1*, *AwayEast2*, *AwayEast1*, *AwayWest-1*, *AwayWest-2*).

2.3.Evaluation of Results

During this final stage, the goodness of fit of each model created in the previous stage will be evaluated. To execute this evaluation, two principal assumptions where the goodness of a linear model are based will be analyzed:

1. **Homogeneity of errors:** The residual errors coming from the model should not follow any particular pattern. They have to be randomly distributed around 0 where higher values may account high errors.

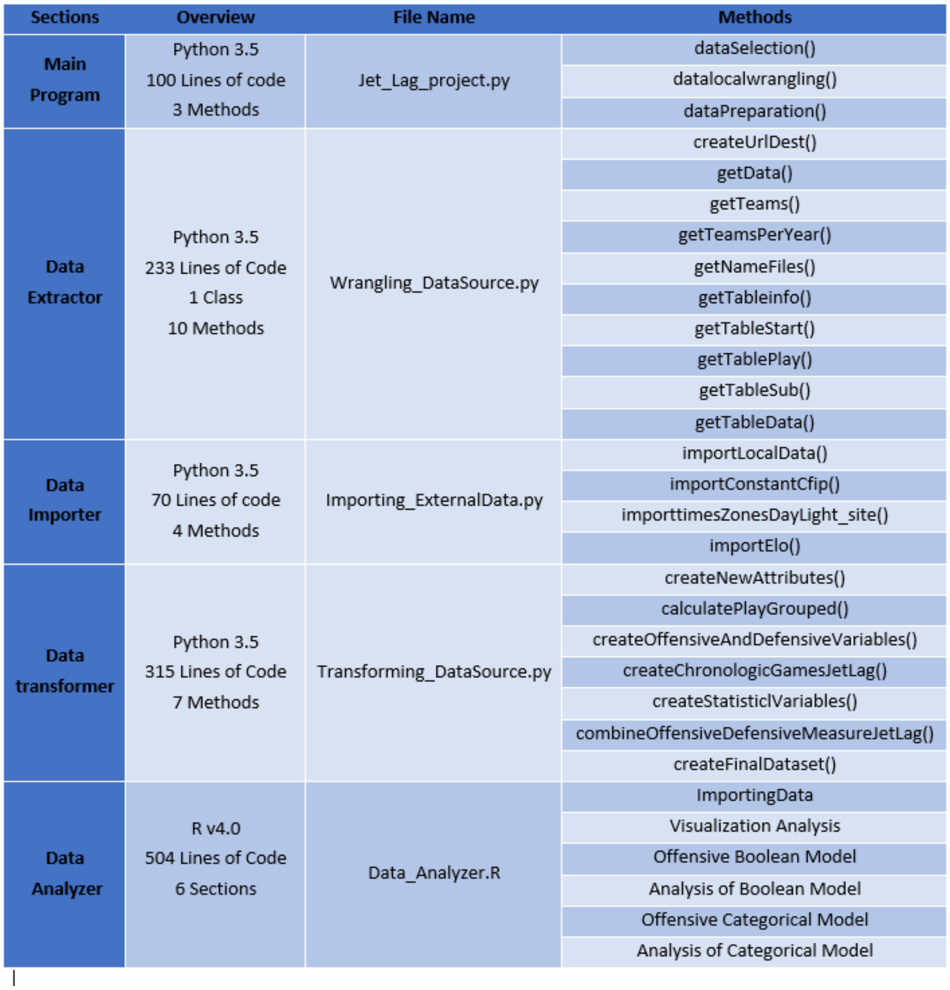
2. **Normality of Errors:** The linear model lies its reliability in the assumption that the errors are independently and normally distributed. Therefore, this behavior needs to be evaluated whether by visual techniques or using statistical testing such as Anderson darling test used to assess normality.

3. Outcomes or Results

The presentation of results has been divided into five general subsections following the methodology presented in the previous numeral. First of all, the overall structure of the code developed (python and R) as the main deliverables for the present project will be presented as outputs for the stage of Analysis and understanding. Then, a breakdown over the results gathered from each section inside the implementation stage will be shown including the outcome from the data selection stage, the transformed data named as data preparation (It includes the pre-processing and transformation stages) and the insights gathered from the data analysis stage. Finally, the evaluation of the results obtained in the previous steps will be summarized and presented.

3.1.Results Analysis and understanding Stage.

The table below shows the principal components including the names of the .py or .R codes and the methods or sections created. A deep explanation about the actions executed during each stage was provided in the methodology section.



Sections	Overview	File Name	Methods

Main Program	Python 3.5
	100 Lines of code
	3 Methods
		Jet_Lag_project.py	dataSelection()
			datalocalwrangling()
			dataPreparation()

Data Extractor	Python 3.5
	233 Lines of Code
	1 Class
	10 Methods
		Wrangling_DataSource.py	createUrlDest()
			getData()
			getTeams()
			getTeamsPerYear()
			getNameFiles()
			getTableinfo()
			getTableStart()
			getTablePlay()
			getTableSub()
			getTableData()

Data Importer	Python 3.5
	70 Lines of code
	4 Methods
		Importing_ExternalData.py	importLocalData()
			importConstantCfip()
			importtimesZonesDayLight_site()
			importElo()

Data transformer	Python 3.5
	315 Lines of Code
	7 Methods
		Transforming_DataSource.py	createNewAttributes()
			calculatePlayGrouped()
			createOffensiveAndDefensiveVariables()
			createChronologicGamesJetLag()
			createStatisticlVariables()
			combineOffensiveDefensiveMeasureJetLag()
			createFinalDataset()

Data Analyzer	R v4.0 504 Lines of Code 6 Sections	Data_Analyzer.R	ImportingData Visualization Analysis Offensive Boolean Model Analysis of Boolean Model Offensive Categorical Model Analysis of Categorical Model

Figure 6. Overall view of methods and files created.

3.2.Results Implementation Stage

The results presented in this section will be based on the analysis of 44 years of data spanning from 1972 to 2018 from the MLB. They encompassed 98266 games for overall 33 different baseball teams.

Data Selection: As was mentioned in the methodology, this part of the program uses the code in *Wrangling_DataSource.py* and *Importing_ExternalData.py* to execute the data extraction from the retrosheet webpage and the loading of local files used by the program. The results of this method are 6 dictionaries where the keys are defined as the years extracted and the values as a dataframe containing the information per each game of each time for the correspondent year. Moreover, 3 dataframes are also produced by this function containing the information related to Elo, Cfip and time zones.

A summary regarding the dictionary outputs of this phase is presented below:

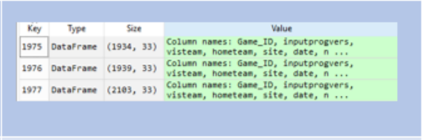
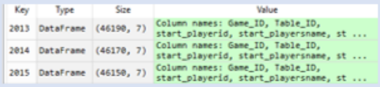
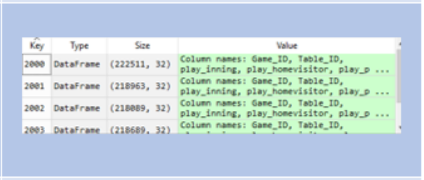
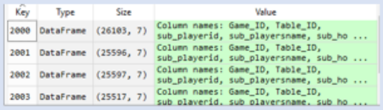
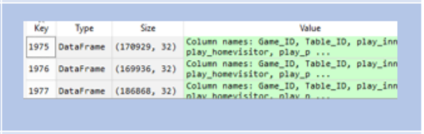
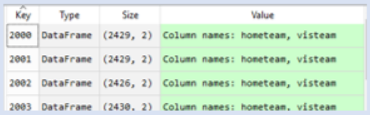
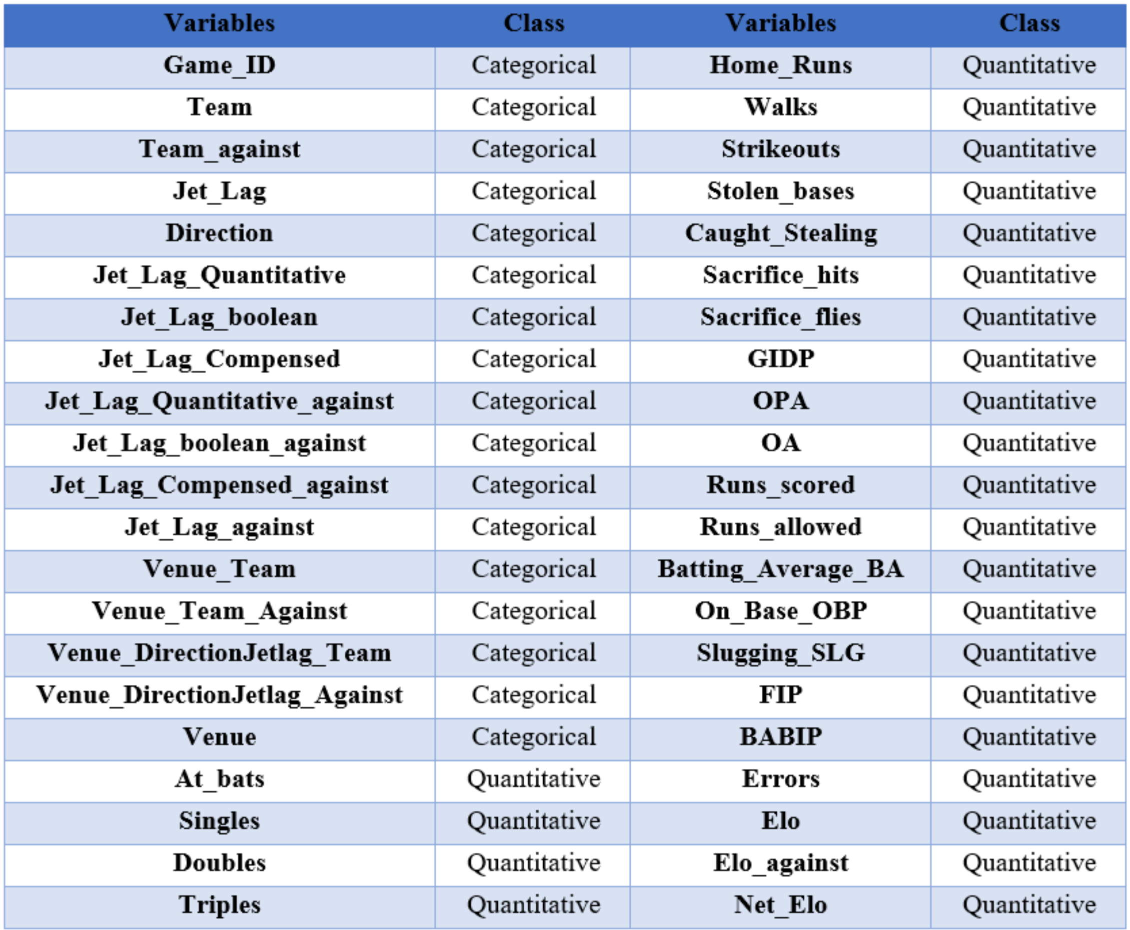
Output	Language	DataType	Dimensions	Evidence
dictInfo*	Python 3.5	Dictionary of Dataframes	(44, ~2100, ~33)	 Key Type Size Value 1975 DataFrame (1934, 33) Column names: Game_ID, inputprogvers, visteam, hometeam, site, date, n ... 1976 DataFrame (1939, 33) Column names: Game_ID, inputprogvers, visteam, hometeam, site, date, n ... 1977 DataFrame (2103, 33) Column names: Game_ID, inputprogvers, visteam, hometeam, site, date, n ...
dictStart*	Python 3.5	Dictionary of Dataframes	(44, ~46170,7)	 Key Type Size Value 2013 DataFrame (46190, 7) Column names: Game_ID, Table_ID, start_playerid, start_playersname, st ... 2014 DataFrame (46170, 7) Column names: Game_ID, Table_ID, start_playerid, start_playersname, st ... 2015 DataFrame (46150, 7) Column names: Game_ID, Table_ID, start_playerid, start_playersname, st ...
dictPlay*	Python 3.5	Dictionary of Dataframes	(44, ~222511, ~33)	 Key Type Size Value 2000 DataFrame (222511, 32) Column names: Game_ID, Table_ID, play_inning, play_homevisitor, play_p ... 2001 DataFrame (218963, 32) Column names: Game_ID, Table_ID, play_inning, play_homevisitor, play_p ... 2002 DataFrame (218089, 32) Column names: Game_ID, Table_ID, play_inning, play_homevisitor, play_p ... 2003 DataFrame (218689, 32) Column names: Game_ID, Table_ID,
dictSub*	Python 3.5	Dictionary of Dataframes	(44,~25596,7)	 Key Type Size Value 2000 DataFrame (26103, 7) Column names: Game_ID, Table_ID, sub_playerid, sub_playersname, sub_ho ... 2001 DataFrame (25596, 7) Column names: Game_ID, Table_ID, sub_playerid, sub_playersname, sub_ho ... 2002 DataFrame (25597, 7) Column names: Game_ID, Table_ID, sub_playerid, sub_playersname, sub_ho ... 2003 DataFrame (25517, 7) Column names: Game_ID, Table_ID, sub_playerid, sub_playersname, sub_ho ...
dictData*	Python 3.5	Dictionary of Dataframes	(44,~170929,~33)	 Key Type Size Value 1975 DataFrame (170929, 32) Column names: Game_ID, Table_ID, play_inn play_homevisitor, play_p ... 1976 DataFrame (169936, 32) Column names: Game_ID, Table_ID, play_inn play_homevisitor, play_p ... 1977 DataFrame (186868, 32) Column names: Game_ID, Table_ID, play_inn play_homevisitor, play_p ...
info_teams*	Python 3.5	Dictionary of Dataframes	(44,~2429,2)	 Key Type Size Value 2000 DataFrame (2429, 2) Column names: hometeam, visteam 2001 DataFrame (2429, 2) Column names: hometeam, visteam 2002 DataFrame (2426, 2) Column names: hometeam, visteam 2003 DataFrame (2430, 2) Column names: hometeam, visteam

Figure 7. Summary of dictionary Outputs.

The last table is created either using the data extractor from the webpage or the local data extractor used from the local files stored in the local repository as was mentioned in the methodology. Finally, the three additional data frames created to represent the ELO, Time zones and Cfip are included in three independent folders inside the zip containing the code. They were created as was mentioned in the methodology and are considered as output artifacts from the Data Selection stage.

Data Preparation: As was mentioned initially, the outcomes of this section combines the outputs coming from data pre-processing and data transformation stages. This part of the program uses the code in Transforming_DataSource.py to create the output. During the process of cleaning, the data was found without any alteration, missing values, outlier or errors which needed any imputation or modification. It was due to the nature of the source which is a legit organization accounting effort over this historical data. Then, as was mentioned in the methodology section, the data provided by the selection stage suffered a series of combinations and transformations which leads to the transformed data, named final dataset as well. This dataset accounted for 196532 rows and 42 columns containing information from the perspective of each team for each game played from 1974 to 2018 in the major league baseball.

A summary regarding the variable names from the output of this phase and also the data type defined to each one is presented below:



Variables	Class	Variables	Class
Game_ID	Categorical	**Home_Runs**	Quantitative
Team	Categorical	**Walks**	Quantitative
Team_against	Categorical	**Strikeouts**	Quantitative
Jet_Lag	Categorical	**Stolen_bases**	Quantitative
Direction	Categorical	**Caught_Stealing**	Quantitative
Jet_Lag_Quantitative	Categorical	**Sacrifice_hits**	Quantitative
Jet_Lag_boolean	Categorical	**Sacrifice_flies**	Quantitative
Jet_Lag_Compensed	Categorical	**GIDP**	Quantitative
Jet_Lag_Quantitative_against	Categorical	**OPA**	Quantitative
Jet_Lag_boolean_against	Categorical	**OA**	Quantitative
Jet_Lag_Compensed_against	Categorical	**Runs_scored**	Quantitative
Jet_Lag_against	Categorical	**Runs_allowed**	Quantitative
Venue_Team	Categorical	**Batting_Average_BA**	Quantitative
Venue_Team_Against	Categorical	**On_Base_OBP**	Quantitative
Venue_DirectionJetlag_Team	Categorical	**Slugging_SLG**	Quantitative
Venue_DirectionJetlag_Against	Categorical	**FIP**	Quantitative
Venue	Categorical	**BABIP**	Quantitative
At_bats	Quantitative	**Errors**	Quantitative
Singles	Quantitative	**Elo**	Quantitative
Doubles	Quantitative	**Elo_against**	Quantitative
Triples	Quantitative	**Net_Elo**	Quantitative

Figure 8. Summary of variable data types.

The nature of Each variable name represented as a column in the dataset is explained through the comments of the code. Moreover, in the previous section of methodology, the objective and function of each variable were also mentioned. Finally, some descriptive statistics over each variable are provided in the appendix A of this document.

Data Analysis: As was mentioned in the methodology, during this stage two different analysis will be executed including an exploratory analysis and a Multivariate Linear Regression analysis. Both analyses were executed in R and the results obtained are coded in the deliverable named *Data_Analyzer.R*

For the exploratory analysis, the general distributions for the main categorical variables will be evaluated, followed by the analysis over the numerical variables. The following images show the general trends showed by the categorical data where some tendencies can be observed.

The figure below analyzes the presence of jetlag by evaluating the jetlag Boolean variable in the dataset. This variable measures the value of Jet Lag after the application of the equation defined in the methodology. It represents the distribution of unique Games-Team instances with any presence of jetlag from the perspective of the principal team.

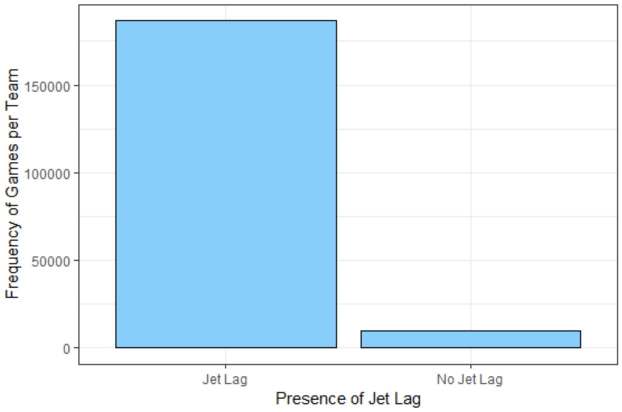


Figure 9. Histogram of Jet lag presence

As is possible to observe, a small percentage of instances were classified as jet lag affected. From the 196532 global instances, only 10016 representing 5.09% were classified as likely candidates to be considered under the effects of jetlag. In the context of the analysis to be developed, only the instances with the presence of jet lag were evaluated. It matches the indications and methodology followed by previous studies in this area (Song, Severini, & Allada, 2017). Then, the data affected by jetlag, about 10016 instances were evaluated. The first visual analysis was to evaluate the balance between the instances of jetlag evaluated from the Venue (Home, Away) and Direction (Eastward or Westward). Here is possible to observe that 41.7% (4185 instances) were classified as eastwards jetlag, whereas 58.3% (5831 instances) are considered as westwards. Moreover, 61.5% of teams playing as Away (6166 instances) and 38.5% (3855 instances) playing as Home were accounting for any presence of jetlag as is shown in the graph below.

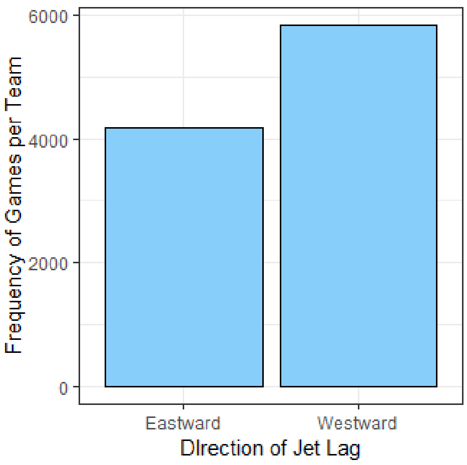


Figure 11. Histogram Direction Jet lag

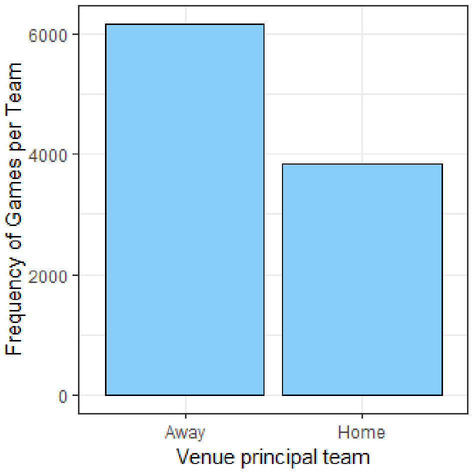


Figure 10. . Histogram Venue Jet lag

From the previous visual representations is possible to evidence that the data under evaluation account a balanced number of instances according to both venue and direction. It allows expecting unbiased results when both variables are evaluated versus the jetlag effects. However, the additional visual analysis should be executed among these variables and the target variables under evaluation. These analyses will be executed in each respective analysis Boolean and categorical.

The following analysis will be executed over the numerical data. The first variable to be analyzed is the ELO, more precisely the net difference among the ELO values for each game-team from the perspective of the principal team. The following image shows the frequency distribution of this variable under analysis.

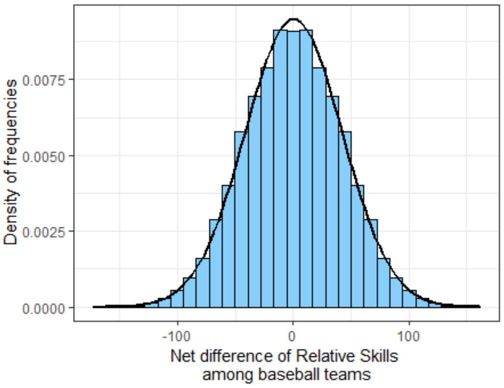


Figure 12. Histogram of Frequencies for Net Elo.

As is possible to observe, this variable follows a normal distribution. It means this variable can be used inside of the multivariate linear regression analysis without considering any additional transformation during the Boolean and categorical analysis.

Before to start the visual analysis correspondent to each experiment either Boolean or categorical, the distribution of frequencies for each offensive statistic should be evaluated. The objective of this analysis is to evaluate the normality in the distribution of each one of the variables. It is particularly important for the success of the experiments due to the nature of the multivariate linear regression analysis which lies in the normal nature of the data. The image below shows the distribution of measurements for each variable:

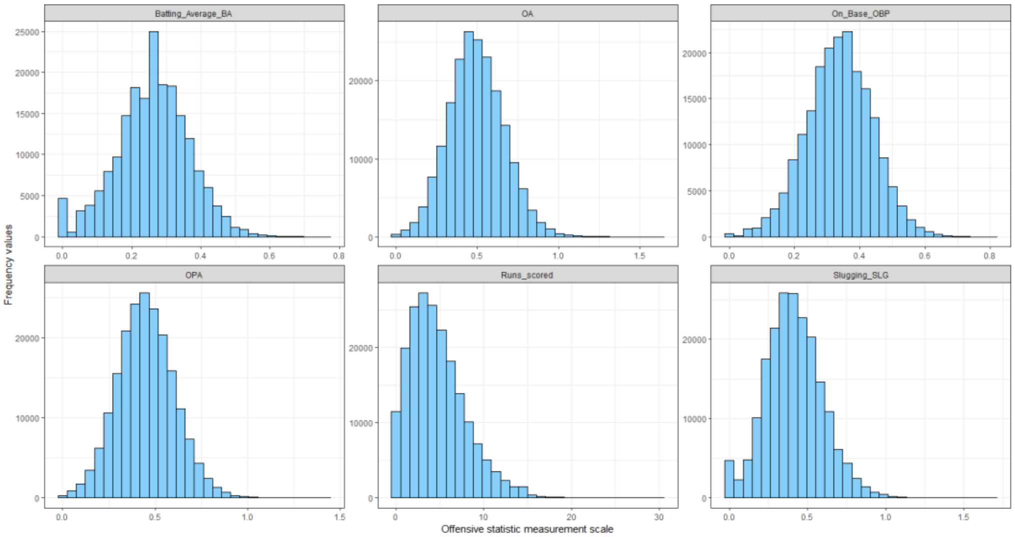


Figure 13. . Histogram of Frequencies for Net Elo by offensive Statistic.

As is possible to observe, the majority offensive statistics show a normal distribution excepting by some slightly positive skewness observed in Run Scored. Therefore, an additional transformation over this variable can be executed. A standard normal transformation is applied resulting in a more normal distributed variable with less positive skewness.

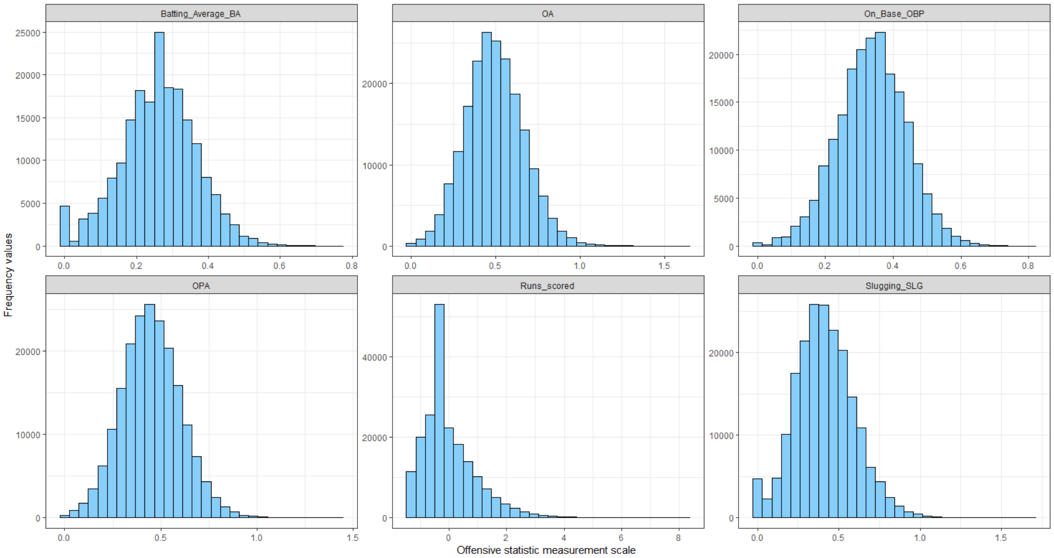


Figure 14. Histogram of Frequencies for Net Elo by offensive statistic transformed.

As was defined in the methodology, the objective of this phase is also to observe the distributions of each offensive statistics versus the jet lag measurement. First of all, the evaluation of jet lag measured as a Boolean variable accounting both direction and venue will be shown. To evaluate this behavior, a visual representation regarding the distribution of each offensive statistic versus the category of Venue + Direction of jetlag will be analyzed.

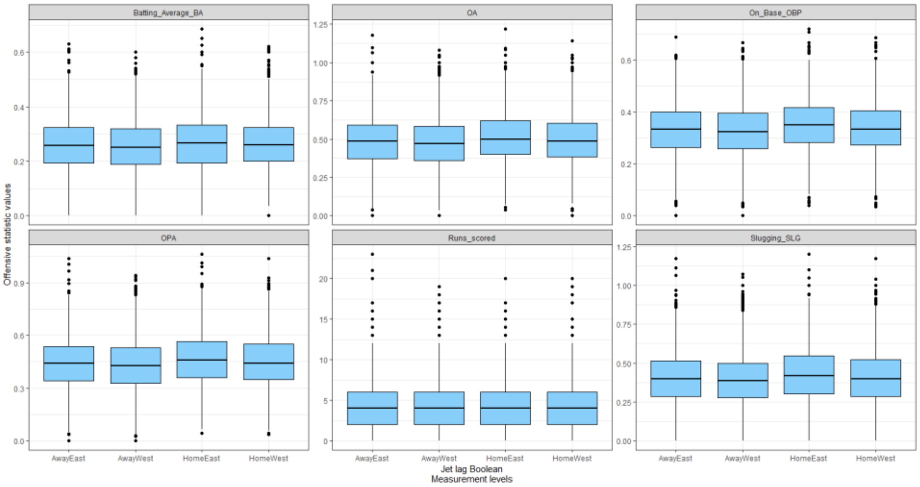


Figure 15. Boxplot distribution for jet lag values by offensive measurements

The previous image allow us to observe that each offensive statistic has a similar distribution among the 4 different combinations of jet lag measured as a Boolean effect. However, is possible to observe slightly variations of the interquartile range (IQR) among each combination of Venue + Direction which can be resulted due to the effect of the value accounted for the jet lag variable. These variations and the significance among the changes observed will be evaluated numerically in the previous section.

On the other hand, the distributions regarding each offensive statistic, the Venue of the team and the jet lag variable measured as a categorical value where the direction is included in the sign is displayed below.

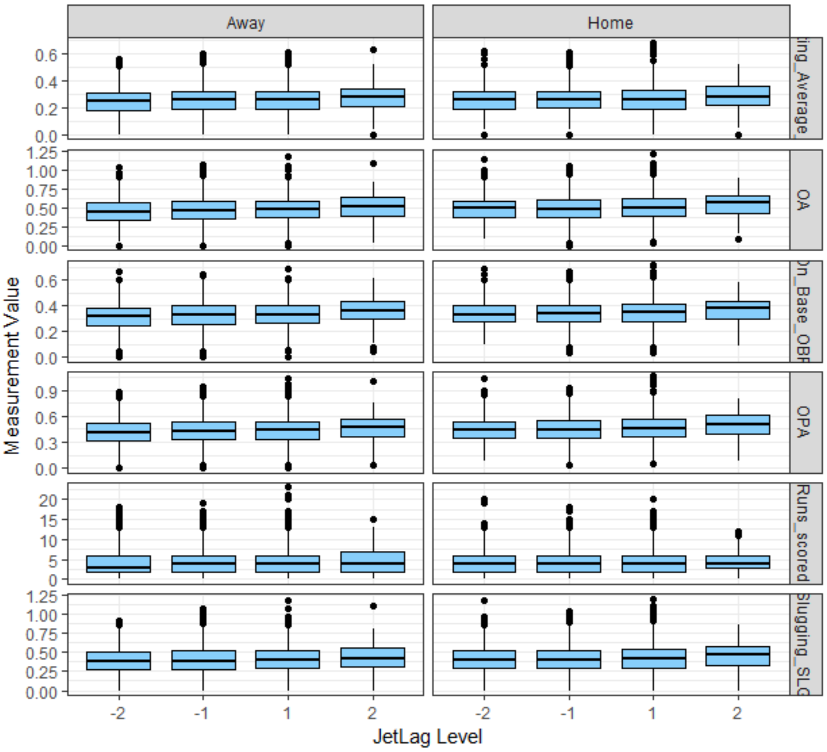


Figure 16. Boxplot distribution for categorical jet lag values by offensive measurements and Home-Away.

The last image shows a similar distribution among all the offensive statistics divided by the categorical variables under evaluation. From the image, is possible to observe slight differences among categories of jet lag in each statistic which appears in both Home and Away teams. These variations and the significance among the changes observed will be evaluated numerically in the previous section.

Findings: From the visualization analysis executed over the dataset extracted, pre-processed and transformed, is possible to evidence:

1. Regarding the direction of jet lag, the data provided is balanced among the number of instances evaluated as jet lag affected for both westwards and Eastwards. This balance is represented by 41.7% of instances eastward and 58.3% of instances westwards.

2. Regarding the venue of each team playing, the data provided is balanced among the number of instances evaluated as jet lag affected for both westwards and Eastwards. This balance is represented by 38.5% instances playing as Home and 61.5% playing as Away.

3. The distribution of the data regarding Elo is normally distributed. It means the data does not show any skewness or kurtosis, therefore, this data may be used to run statistical tests where normality is assumed.

4. The distribution of the data regarding the values of each offensive statistic is normally distributed. It means the data does not show any skewness or kurtosis, therefore, this data may be used to run statistical tests where normality is assumed.

5. The distribution of each jet lag related variable versus each offensive statistic seems to have a similar spread accounting with slight variations among categories which should be further analyzed.

After observing some slight differences among the measured values for the six offensive statistics under evaluation, is necessary to evaluate numerically the level of significance among those variations. To capture the effect of each one of the variables as a whole, the approach implemented by previous studies was followed as was shown in the methodology (Song, Severini, & Allada, 2017). Therefore, multivariate linear regression analysis for each offensive statistic was executed. This analysis was executed for both Boolean and categorical representation of jet lag.

Boolean Analysis.

The first numerical analysis to be executed is the multivariate linear regression analysis for the Boolean approach. As was described in the methodology, four variables representing the presence or absence of jet lag as a team playing whether home or Away and with the effect whether towards east or west are included in the equation. Moreover, the teams playing component and the Net value of ELO are also considering.

The final equation for the linear regression analysis in the Boolean approach is:

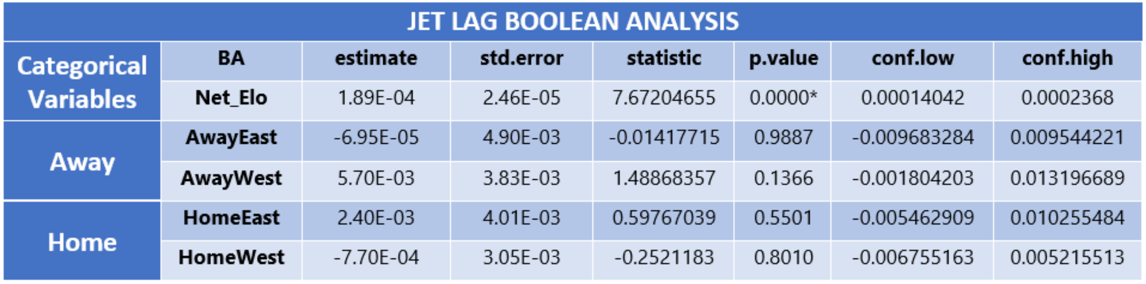
$$\begin{aligned}
Statistic = \ & \delta_1 HomeWest + \delta_2 HomeEast + \delta_3 AwayWest + \delta_4 AwayEast + \delta_5 AwayANA \\
& + \delta_6 AwayARI + \delta_7 AwayATL + \delta_8 AwayBAL + \delta_9 AwayBOS + \delta_{10} AwayCAL \\
& + \delta_{11} AwayCHA + \delta_{12} AwayCHN + \delta_{13} AwayCIN + \delta_{14} AwayCLE \\
& + \delta_{15} AwayCOL + \delta_{16} AwayDET + \delta_{17} AwayFLO + \delta_{18} AwayHOU \\
& + \delta_{19} AwayKCA + \delta_{20} AwayLAN + \delta_{21} AwayMIA + \delta_{22} AwayMIL \\
& + \delta_{23} AwayMIN + \delta_{24} AwayMON + \delta_{25} AwayNYA + \delta_{26} AwayNYN \\
& + \delta_{27} AwayOAK + \delta_{28} AwayPHI + \delta_{29} AwayPIT + \delta_{30} AwaySDN + \delta_{31} AwaySEA \\
& + \delta_{32} AwaySFN + \delta_{33} AwaySLN + \delta_{34} AwayTBA + \delta_{35} AwayTEX \\
& + \delta_{36} AwayTOR + \delta_{37} AwayWAS + \delta_{38} HomeANA + \delta_{39} HomeARI \\
& + \delta_{40} HomeATL + \delta_{41} HomeBAL + \delta_{42} HomeBOS + \delta_{43} HomeCAL \\
& + \delta_{44} HomeCHA + \delta_{45} HomeCHN + \delta_{46} HomeCIN + \delta_{47} HomeCLE \\
& + \delta_{48} HomeCOL + \delta_{49} HomeDET + \delta_{50} HomeFLO + \delta_{51} HomeHOU \\
& + \delta_{52} HomeKCA + \delta_{53} HomeLAN + \delta_{54} HomeMIA + \delta_{55} HomeMIL \\
& + \delta_{56} HomeMIN + \delta_{57} HomeMON + \delta_{58} HomeNYA + \delta_{59} HomeNYN \\
& + \delta_{60} HomeOAK + \delta_{61} HomePHI + \delta_{62} HomePIT + \delta_{63} HomeSDN \\
& + \delta_{64} HomeSEA + \delta_{65} HomeSFN + \delta_{66} HomeSLN + \delta_{67} HomeTBA \\
& + \delta_{68} HomeTEX + \delta_{69} HomeTOR + \delta_{70} HomeWAS + \delta_{71} Elo
\end{aligned}$$

The following tables show the results obtained for each coefficient after applying the multivariate linear regression over each offensive statistic. Only the more relevant variables under evaluation were selected to be displayed as results including the values for $\delta_1, \delta_2, \delta_3, \delta_4\ and\ \delta_{71}$. Moreover, is important to emphasize that each coefficient shown below contain the effect of the other variables over its importance. Moreover, the standard error of the coefficient, the significance level (p value) and the confidence intervals for each parameter are included in the table.

As was mentioned in the methodology, the objective of the present analysis is to evaluate the level of significance of each parameter by applying a t-test evaluation (2 tails or 1 tail depending of the variable under evaluation) where the null hypothesis is considered as the coefficient associated to the parameter is equal to 0. On the other hand, the alternative hypothesis is stated as the value is effectively different to zero and therefore significant. For the sake of these experiments, a level of significance of $\alpha = 5\%$ was defined for each offensive statistic.

The following tables show the main parameters extracted after creating the linear model for each offensive statistic, also each statistic will be accompanied by its respective results.

1. *Batting Average*

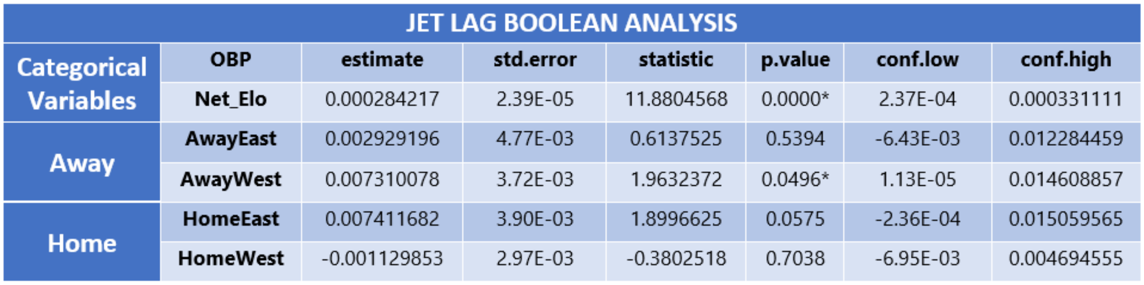


JET LAG BOOLEAN ANALYSIS							
Categorical Variables	BA	estimate	std.error	statistic	p.value	conf.low	conf.high
	Net_Elo	1.89E-04	2.46E-05	7.67204655	0.0000*	0.00014042	0.0002368
Away	AwayEast	-6.95E-05	4.90E-03	-0.01417715	0.9887	-0.009683284	0.009544221
	AwayWest	5.70E-03	3.83E-03	1.48868357	0.1366	-0.001804203	0.013196689
Home	HomeEast	2.40E-03	4.01E-03	0.59767039	0.5501	-0.005462909	0.010255484
	HomeWest	-7.70E-04	3.05E-03	-0.2521183	0.8010	-0.006755163	0.005215513

The results for *Batting average* show the following significant variables:

- $NetElo = \ 0.000189 \pm 0.0000246$

2. *On Base performance*

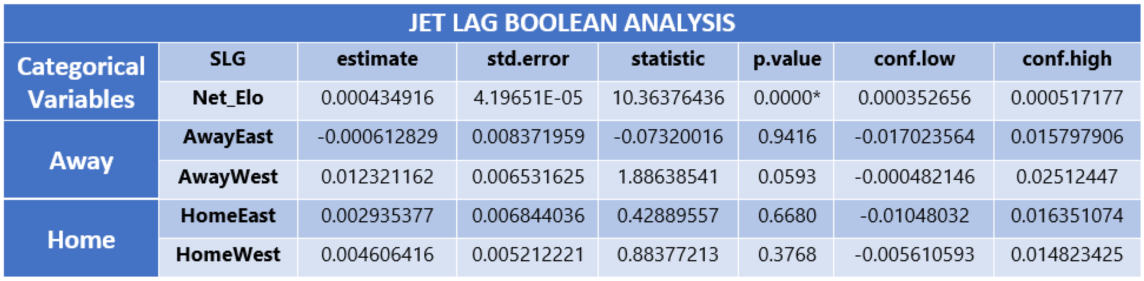


JET LAG BOOLEAN ANALYSIS							
Categorical Variables	OBP	estimate	std.error	statistic	p.value	conf.low	conf.high
	Net_Elo	0.000284217	2.39E-05	11.8804568	0.0000*	2.37E-04	0.000331111
Away	AwayEast	0.002929196	4.77E-03	0.6137525	0.5394	-6.43E-03	0.012284459
	AwayWest	0.007310078	3.72E-03	1.9632372	0.0496*	1.13E-05	0.014608857
Home	HomeEast	0.007411682	3.90E-03	1.8996625	0.0575	-2.36E-04	0.015059565
	HomeWest	-0.001129853	2.97E-03	-0.3802518	0.7038	-6.95E-03	0.004694555

The results for *On Base Performance* show the following significant variables:

- $NetElo = 0.000284217 \pm 0.0000239$
- $AwayWest = 0.0007310078 \pm 0.00372$

3. *Slugging Percentage*

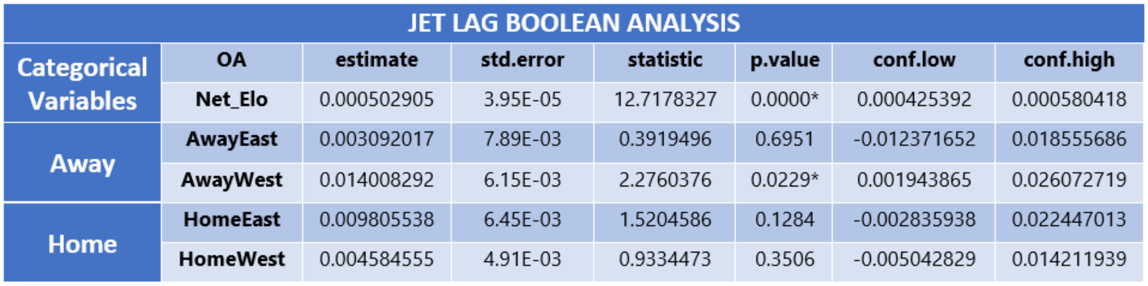


JET LAG BOOLEAN ANALYSIS							
Categorical Variables	SLG	estimate	std.error	statistic	p.value	conf.low	conf.high
	Net_Elo	0.000434916	4.19651E-05	10.36376436	0.0000*	0.000352656	0.000517177
Away	AwayEast	-0.000612829	0.008371959	-0.07320016	0.9416	-0.017023564	0.015797906
	AwayWest	0.012321162	0.006531625	1.88638541	0.0593	-0.000482146	0.02512447
Home	HomeEast	0.002935377	0.006844036	0.42889557	0.6680	-0.01048032	0.016351074
	HomeWest	0.004606416	0.005212221	0.88377213	0.3768	-0.005610593	0.014823425

The results for *Slugging Percentage* show the following significant variables:

- $NetElo = 0.000434916 \pm 0.0000419$

4. *Offensive Average*

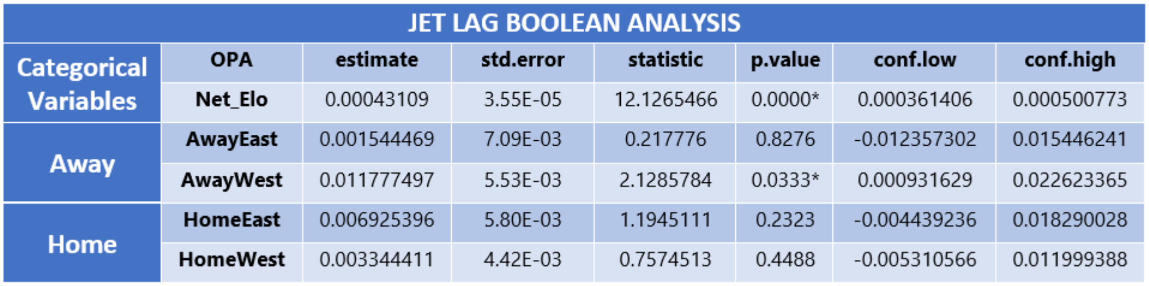


JET LAG BOOLEAN ANALYSIS							
Categorical Variables	OA	estimate	std.error	statistic	p.value	conf.low	conf.high
	Net_Elo	0.000502905	3.95E-05	12.7178327	0.0000*	0.000425392	0.000580418
Away	AwayEast	0.003092017	7.89E-03	0.3919496	0.6951	-0.012371652	0.018555686
	AwayWest	0.014008292	6.15E-03	2.2760376	0.0229*	0.001943865	0.026072719
Home	HomeEast	0.009805538	6.45E-03	1.5204586	0.1284	-0.002835938	0.022447013
	HomeWest	0.004584555	4.91E-03	0.9334473	0.3506	-0.005042829	0.014211939

The results for *Offensive Average* show the following significant variables:

- $NetElo = 0.000502905 \pm 0.0000395$
- $AwayWest = 0.014008292 \pm 0.00615$

5. *Offensive Performance Average:*



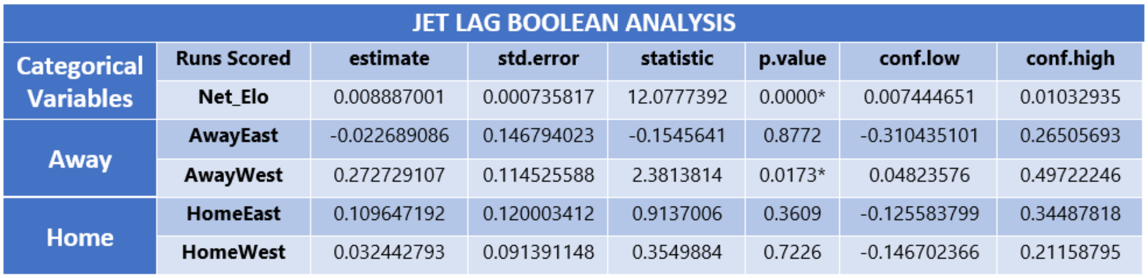
JET LAG BOOLEAN ANALYSIS

Categorical Variables	OPA	estimate	std.error	statistic	p.value	conf.low	conf.high
	Net_Elo	0.00043109	3.55E-05	12.1265466	0.0000*	0.000361406	0.000500773
Away	AwayEast	0.001544469	7.09E-03	0.217776	0.8276	-0.012357302	0.015446241
	AwayWest	0.011777497	5.53E-03	2.1285784	0.0333*	0.000931629	0.022623365
Home	HomeEast	0.006925396	5.80E-03	1.1945111	0.2323	-0.004439236	0.018290028
	HomeWest	0.003344411	4.42E-03	0.7574513	0.4488	-0.005310566	0.011999388

The results for *Offensive Performance Average* show the following significant variables:

- $NetElo = \ 0.00043109 \pm 0.0000355$
- $AwayWest = \ 0.011777497 \pm 0.00553$

6. *Runs Scored*



JET LAG BOOLEAN ANALYSIS

Categorical Variables	Runs Scored	estimate	std.error	statistic	p.value	conf.low	conf.high
	Net_Elo	0.008887001	0.000735817	12.0777392	0.0000*	0.007444651	0.01032935
Away	AwayEast	-0.022689086	0.146794023	-0.1545641	0.8772	-0.310435101	0.26505693
	AwayWest	0.272729107	0.114525588	2.3813814	0.0173*	0.04823576	0.49722246
Home	HomeEast	0.109647192	0.120003412	0.9137006	0.3609	-0.125583799	0.34487818
	HomeWest	0.032442793	0.091391148	0.3549884	0.7226	-0.146702366	0.21158795

The results for *Runs Scored* show the following significant variables:

- $NetElo = \ 0.00888701 \pm 0.007358$
- $AwayWest = \ 0.272729107 \pm 0.114525588$

Findings: Regarding the multivariate linear regression analysis for the Boolean approach, is possible to evidence that:

1. The Net Elo variables included in the analysis are highly correlated to the values gathered from all the offensive statistics. This relation shows a positive relationship which infers that when the difference of relative skills is higher from the principal team perspective, the offensive statistic is higher.

2. There is a clear predominance of the venue Away over the venue Home in the results. It might infer that the effects when the jetlag is faced playing as away team can be more significant than the same effects playing as a home team.

3. There is a clear predominance of the jetlag faced westwards over eastwards in the results. It might infer that the effects of jetlag over offensive statistics of teams who travelled westwards can be more significant than the effects over teams travelling eastwards.

Categorical Analysis.
The second numerical analysis to be executed is the multivariate linear regression analysis for the Categorical approach. As was described in the methodology, nine variables representing the levels (-2,-1,1,2) of jet lag as a team playing whether home or Away and with the effect whether towards

east or west are included in the equation. Moreover, the teams playing component and the Net value of ELO are also considering.

The final equation for the linear regression analysis in the Boolean approach is:

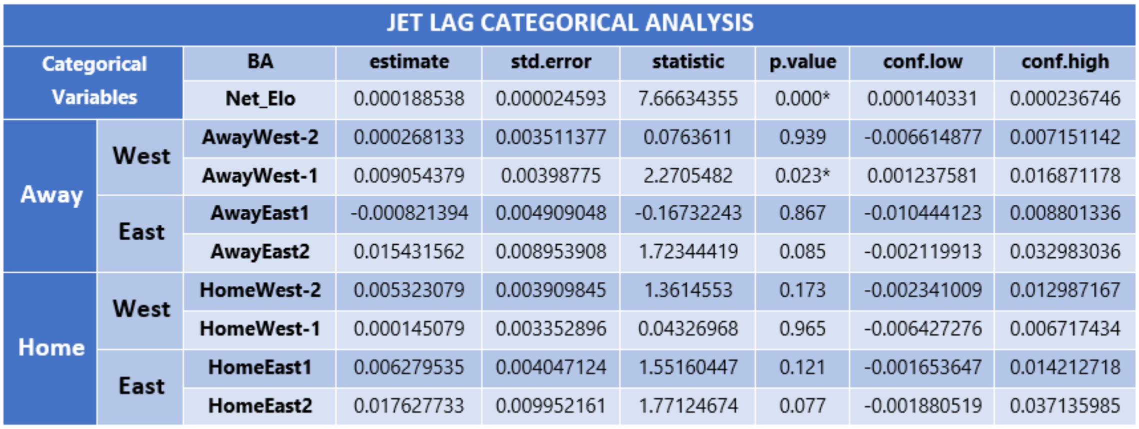
$$\begin{aligned}
Statistic = {} & \delta_1 AwayEast2 + \delta_2 AwayEast1 + \delta_3 AwayWest - 2 + \delta_4 AwayEast - 1 \\
& + \delta_5 AwayANA + \delta_6 AwayARI + \delta_7 AwayATL + \delta_8 AwayBAL + \delta_9 AwayBOS \\
& + \delta_{10} AwayCAL + \delta_{11} AwayCHA + \delta_{12} AwayCHN + \delta_{13} AwayCIN \\
& + \delta_{14} AwayCLE + \delta_{15} AwayCOL + \delta_{16} AwayDET + \delta_{17} AwayFLO \\
& + \delta_{18} AwayHOU + \delta_{19} AwayKCA + \delta_{20} AwayLAN + \delta_{21} AwayMIA \\
& + \delta_{22} AwayMIL + \delta_{23} AwayMIN + \delta_{24} AwayMON + \delta_{25} AwayNYA \\
& + \delta_{26} AwayNYN + \delta_{27} AwayOAK + \delta_{28} AwayPHI + \delta_{29} AwayPIT \\
& + \delta_{30} AwaySDN + \delta_{31} AwaySEA + \delta_{32} AwaySFN + \delta_{33} AwaySLN \\
& + \delta_{34} AwayTBA + \delta_{35} AwayTEX + \delta_{36} AwayTOR + \delta_{37} AwayWAS \\
& + \delta_{38} HomeANA + \delta_{39} HomeARI + \delta_{40} HomeATL + \delta_{41} HomeBAL \\
& + \delta_{42} HomeBOS + \delta_{43} HomeCAL + \delta_{44} HomeCHA + \delta_{45} HomeCHN \\
& + \delta_{46} HomeCIN + \delta_{47} HomeCLE + \delta_{48} HomeCOL + \delta_{49} HomeDET \\
& + \delta_{50} HomeFLO + \delta_{51} HomeHOU + \delta_{52} HomeKCA + \delta_{53} HomeLAN \\
& + \delta_{54} HomeMIA + \delta_{55} HomeMIL + \delta_{56} HomeMIN + \delta_{57} HomeMON \\
& + \delta_{58} HomeNYA + \delta_{59} HomeNYN + \delta_{60} HomeOAK + \delta_{61} HomePHI \\
& + \delta_{62} HomePIT + \delta_{63} HomeSDN + \delta_{64} HomeSEA + \delta_{65} HomeSFN \\
& + \delta_{66} HomeSLN + \delta_{67} HomeTBA + \delta_{68} HomeTEX + \delta_{69} HomeTOR \\
& + \delta_{70} HomeWAS + \delta_{71} HomeEast2 + \delta_{72} HomeEast1 + \delta_{73} HomeWest - 1 \\
& + \delta_{74} HomeWest - 2 + \delta_{75} Elo
\end{aligned}$$

The following tables show the results obtained for each coefficient after applying the multivariate linear regression over each offensive statistic. Only the more relevant variables under evaluation were selected to be displayed as results, including the values for $\delta_1, \delta_2, \delta_3, \delta_4, \delta_{71}, \delta_{72}, \delta_{73}, \delta_{74}\ and\ \delta_{75}$. Moreover, is important to emphasize that each coefficient shown below contain the effect of the other variables over its importance. Moreover, the standard error of the coefficient, the significance level (p value) and the confidence intervals for each parameter are included in the table.

As was mentioned in the methodology, the objective of the present analysis is to evaluate the level of significance of each parameter by applying a t-test evaluation (2 tails or 1 tail depending on the variable under evaluation) where the null hypothesis is considered as the coefficient associated to the parameter is equal to 0. On the other hand, the alternative hypothesis is stated as the value is effectively different to zero and therefore significant. For the sake of these experiments, a level of significance of $\alpha = 5\%$ was defined for each offensive statistic.

The following tables show the main parameters extracted after creating the linear model for each offensive statistic, also each statistic will be accompanied by its respective results.

1. *Batting Average*



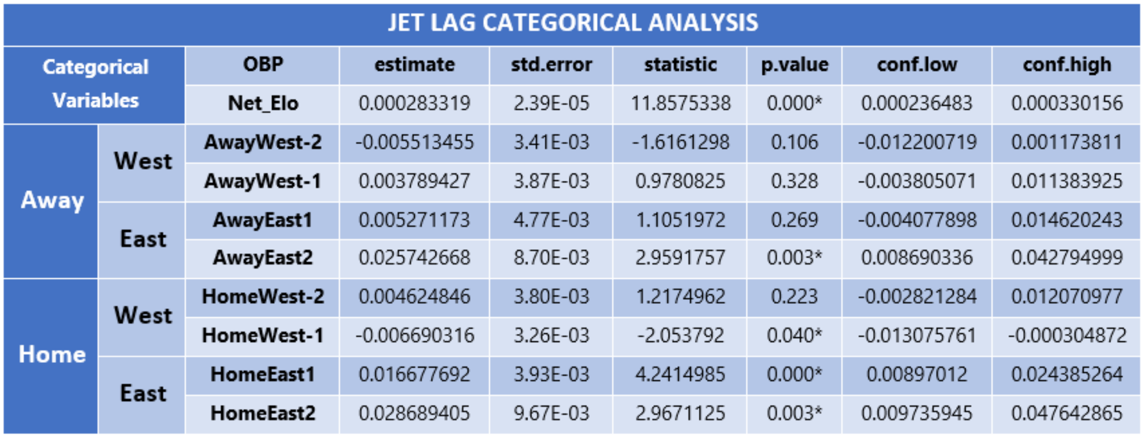
JET LAG CATEGORICAL ANALYSIS

Categorical Variables		BA	estimate	std.error	statistic	p.value	conf.low	conf.high
		Net_Elo	0.000188538	0.000024593	7.66634355	0.000*	0.000140331	0.000236746
Away	West	AwayWest-2	0.000268133	0.003511377	0.0763611	0.939	-0.006614877	0.007151142
		AwayWest-1	0.009054379	0.00398775	2.2705482	0.023*	0.001237581	0.016871178
	East	AwayEast1	-0.000821394	0.004909048	-0.16732243	0.867	-0.010444123	0.008801336
		AwayEast2	0.015431562	0.008953908	1.72344419	0.085	-0.002119913	0.032983036
Home	West	HomeWest-2	0.005323079	0.003909845	1.3614553	0.173	-0.002341009	0.012987167
		HomeWest-1	0.000145079	0.003352896	0.04326968	0.965	-0.006427276	0.006717434
	East	HomeEast1	0.006279535	0.004047124	1.55160447	0.121	-0.001653647	0.014212718
		HomeEast2	0.017627733	0.009952161	1.77124674	0.077	-0.001880519	0.037135985

The results for *Batting Average* show the following significant variables:

- ○ $NetElo = 0.000188538 \pm 0.000024593$
- ○ $AwayWest_{-1} = 0.009054379 \pm 0.00398775$

2. *On Base Performance*



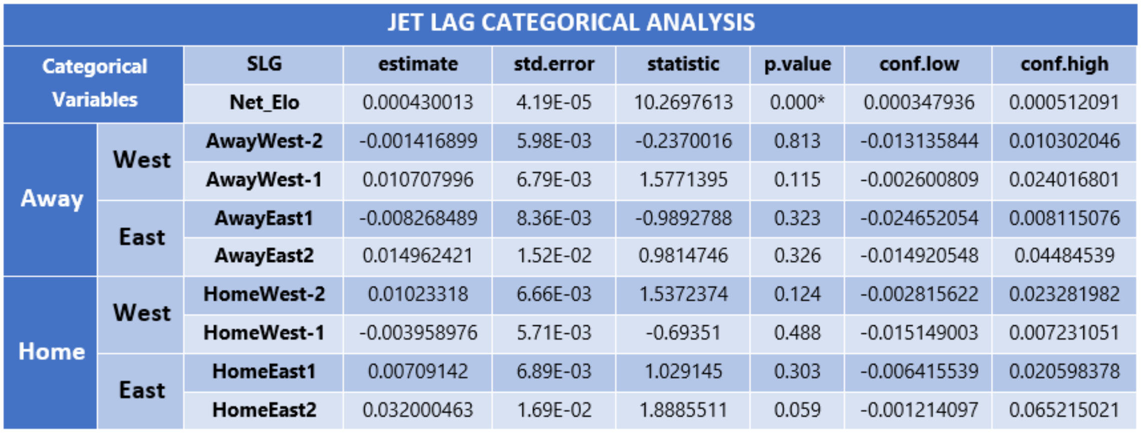
JET LAG CATEGORICAL ANALYSIS

Categorical Variables		OBP	estimate	std.error	statistic	p.value	conf.low	conf.high
		Net_Elo	0.000283319	2.39E-05	11.8575338	0.000*	0.000236483	0.000330156
Away	West	AwayWest-2	-0.005513455	3.41E-03	-1.6161298	0.106	-0.012200719	0.001173811
		AwayWest-1	0.003789427	3.87E-03	0.9780825	0.328	-0.003805071	0.011383925
	East	AwayEast1	0.005271173	4.77E-03	1.1051972	0.269	-0.004077898	0.014620243
		AwayEast2	0.025742668	8.70E-03	2.9591757	0.003*	0.008690336	0.042794999
Home	West	HomeWest-2	0.004624846	3.80E-03	1.2174962	0.223	-0.002821284	0.012070977
		HomeWest-1	-0.006690316	3.26E-03	-2.053792	0.040*	-0.013075761	-0.000304872
	East	HomeEast1	0.016677692	3.93E-03	4.2414985	0.000*	0.00897012	0.024385264
		HomeEast2	0.028689405	9.67E-03	2.9671125	0.003*	0.009735945	0.047642865

The results for *On Base Performance* show the following significant variables:

- ○ $NetElo = 0.000283319 \pm 0.0000239$
- ○ $AwayEast_2 = 0.025742668 \pm 0.0087$
- ○ $HomeWest_{-1} = -0.006699 \pm 0.00398775$
- ○ $HomeEast_1 = 0.016677692 \pm 0.00393$
- ○ $HomeEast_2 = 0.028689405 \pm 0.00967$

3. *Slugging Percentage*



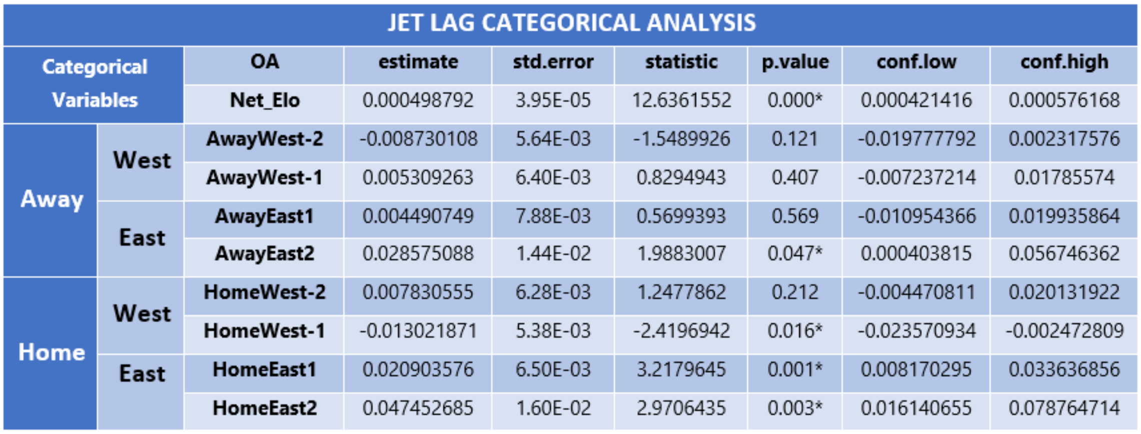
JET LAG CATEGORICAL ANALYSIS

Categorical Variables		SLG	estimate	std.error	statistic	p.value	conf.low	conf.high
		Net_Elo	0.000430013	4.19E-05	10.2697613	0.000*	0.000347936	0.000512091
Away	West	AwayWest-2	-0.001416899	5.98E-03	-0.2370016	0.813	-0.013135844	0.010302046
		AwayWest-1	0.010707996	6.79E-03	1.5771395	0.115	-0.002600809	0.024016801
	East	AwayEast1	-0.008268489	8.36E-03	-0.9892788	0.323	-0.024652054	0.008115076
		AwayEast2	0.014962421	1.52E-02	0.9814746	0.326	-0.014920548	0.04484539
Home	West	HomeWest-2	0.01023318	6.66E-03	1.5372374	0.124	-0.002815622	0.023281982
		HomeWest-1	-0.003958976	5.71E-03	-0.69351	0.488	-0.015149003	0.007231051
	East	HomeEast1	0.00709142	6.89E-03	1.029145	0.303	-0.006415539	0.020598378
		HomeEast2	0.032000463	1.69E-02	1.8885511	0.059	-0.001214097	0.065215021

The results for *Slugging Percentage* show the following significant variables:

- ○ $NetElo = 0.000430013 \pm 0.000024593$

4. *Offensive Average*



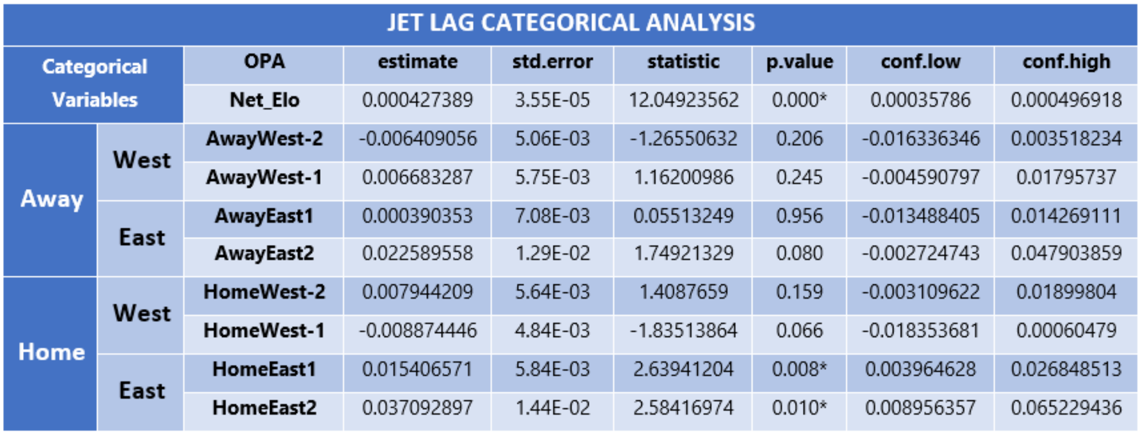
JET LAG CATEGORICAL ANALYSIS

Categorical Variables		OA	estimate	std.error	statistic	p.value	conf.low	conf.high
		Net_Elo	0.000498792	3.95E-05	12.6361552	0.000*	0.000421416	0.000576168
Away	West	AwayWest-2	-0.008730108	5.64E-03	-1.5489926	0.121	-0.019777792	0.002317576
		AwayWest-1	0.005309263	6.40E-03	0.8294943	0.407	-0.007237214	0.01785574
	East	AwayEast1	0.004490749	7.88E-03	0.5699393	0.569	-0.010954366	0.019935864
		AwayEast2	0.028575088	1.44E-02	1.9883007	0.047*	0.000403815	0.056746362
Home	West	HomeWest-2	0.007830555	6.28E-03	1.2477862	0.212	-0.004470811	0.020131922
		HomeWest-1	-0.013021871	5.38E-03	-2.4196942	0.016*	-0.023570934	-0.002472809
	East	HomeEast1	0.020903576	6.50E-03	3.2179645	0.001*	0.008170295	0.033636856
		HomeEast2	0.047452685	1.60E-02	2.9706435	0.003*	0.016140655	0.078764714

The results for *Offensive Average* show the following significant variables:

- ○ $NetElo = 0.000498792 \pm 0.0000395$
- ○ $AwayEast_2 = 0.028575088 \pm 0.0144$
- ○ $HomeWest_{-1} = 0.007830555 \pm 0.00628$
- ○ $HomeEast_1 = 0.020903576 \pm 0.0065$
- ○ $HomeEast_2 = 0.047452685 \pm 0.0160$

5. *Offensive Performance Average*

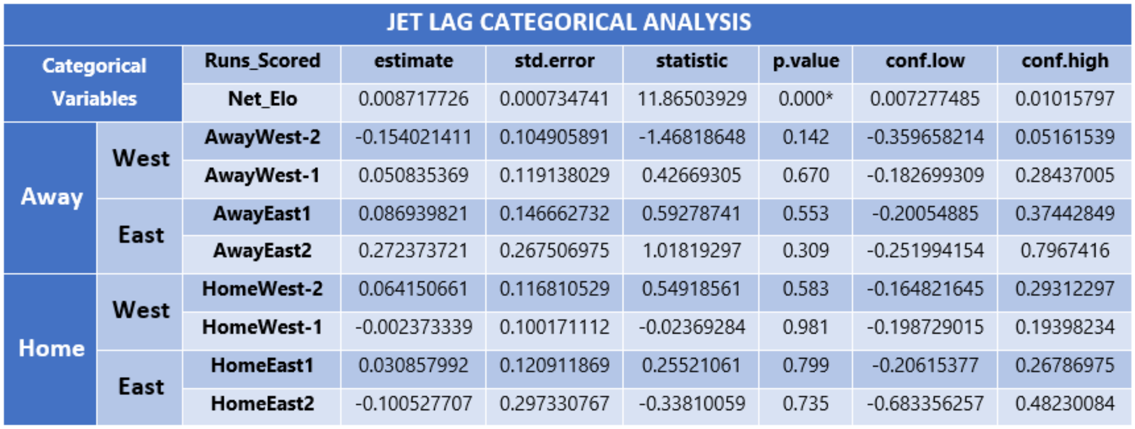


JET LAG CATEGORICAL ANALYSIS								
Categorical Variables		OPA	estimate	std.error	statistic	p.value	conf.low	conf.high
		Net_Elo	0.000427389	3.55E-05	12.04923562	0.000*	0.00035786	0.000496918
Away	West	AwayWest-2	-0.006409056	5.06E-03	-1.26550632	0.206	-0.016336346	0.003518234
		AwayWest-1	0.006683287	5.75E-03	1.16200986	0.245	-0.004590797	0.01795737
	East	AwayEast1	0.000390353	7.08E-03	0.05513249	0.956	-0.013488405	0.014269111
		AwayEast2	0.022589558	1.29E-02	1.74921329	0.080	-0.002724743	0.047903859
Home	West	HomeWest-2	0.007944209	5.64E-03	1.4087659	0.159	-0.003109622	0.01899804
		HomeWest-1	-0.008874446	4.84E-03	-1.83513864	0.066	-0.018353681	0.00060479
	East	HomeEast1	0.015406571	5.84E-03	2.63941204	0.008*	0.003964628	0.026848513
		HomeEast2	0.037092897	1.44E-02	2.58416974	0.010*	0.008956357	0.065229436

The results for *Offensive Average* show the following significant variables:

- $NetElo = 0.000427389 \pm 0.0000355$
- $HomeEast_1 = 0.015406571 \pm 0.00584$
- $HomeEast_2 = 0.037092897 \pm 0.0144$

6. *Runs Scored*



JET LAG CATEGORICAL ANALYSIS								
Categorical Variables		Runs_Scored	estimate	std.error	statistic	p.value	conf.low	conf.high
		Net_Elo	0.008717726	0.000734741	11.86503929	0.000*	0.007277485	0.01015797
Away	West	AwayWest-2	-0.154021411	0.104905891	-1.46818648	0.142	-0.359658214	0.05161539
		AwayWest-1	0.050835369	0.119138029	0.42669305	0.670	-0.182699309	0.28437005
	East	AwayEast1	0.086939821	0.146662732	0.59278741	0.553	-0.20054885	0.37442849
		AwayEast2	0.272373721	0.267506975	1.01819297	0.309	-0.251994154	0.7967416
Home	West	HomeWest-2	0.064150661	0.116810529	0.54918561	0.583	-0.164821645	0.29312297
		HomeWest-1	-0.002373339	0.100171112	-0.02369284	0.981	-0.198729015	0.19398234
	East	HomeEast1	0.030857992	0.120911869	0.25521061	0.799	-0.20615377	0.26786975
		HomeEast2	-0.100527707	0.297330767	-0.33810059	0.735	-0.683356257	0.48230084

The results for *Runs Scored* show the following significant variables:

- $NetElo = 0.008717726 \pm 0.000734741$

Findings: Regarding the multivariate linear regression analysis for the Categorical approach, is possible to evidence that:

1. The Net Elo variables included in the analysis are highly correlated to the values gathered from all the offensive statistics. This relation shows a positive relationship which infers that when the difference of relative skills is higher from the principal team perspective, the offensive statistic is higher.

2. There is a clear predominance of the venue Home over the venue Away in the results. It might infer that the effects when the jetlag is faced playing as Home team can be more significant than the same effects playing as an Away team.

3. There is not a clear predominance of the jetlag faced westwards over eastwards in the results. It might infer that the effects of jetlag over offensive statistics of teams when they are evaluated as independent categories do not show any significant difference. However, in terms of the number of significant values gathered, the effects of travelling eastwards look more significant than travelling westwards.

3.3.Results Evaluation Stage

As was explained in the methodology, this stage focused on the assumption evaluation of the two linear models implemented previously. Each assessment will be executed graphically and where be necessary, a numerical analysis will be included to support the result gathered. The results over the model created for the offensive statistic of Offensive performance Average (OPA) is shown in this section. The other statistics followed the same behavior but were not included in the report. The analysis will be executed independently for each linear model implemented:

1. *Boolean Linear Model*

a. *Homogeneity of errors:*
The figure below assess the Homogeneity of errors in the linear model from OPA. It shows the distribution of errors across the fitted values for the linear model:

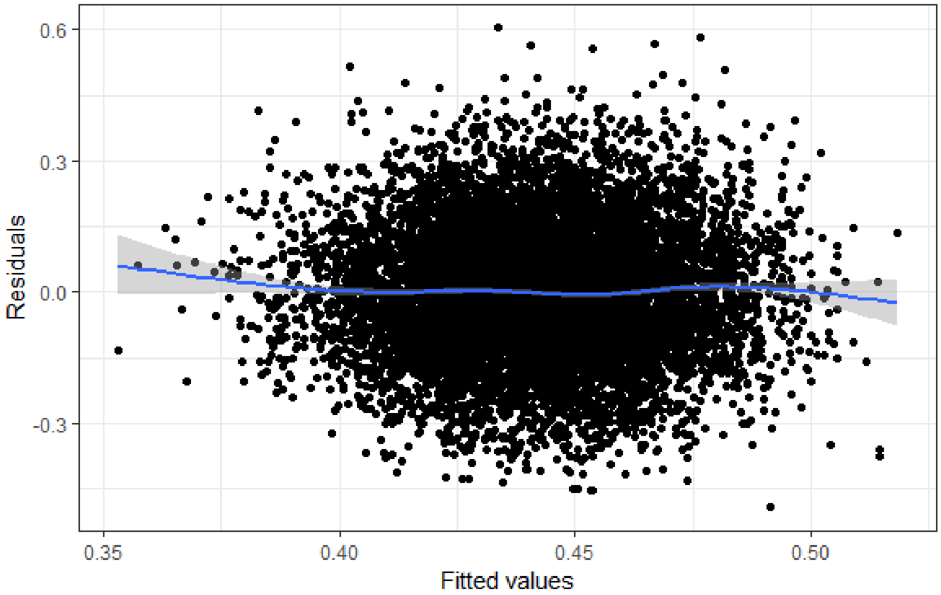


Figure 17. Homogeneity of Errors Boolean Model

The figure below assess the Homogeneity of errors across the different levels of measurement for the jet lag presence in the Boolean linear model from OPA. It shows the distribution of errors across the fitted values for the linear model in each one of the main categories under evaluation for the Boolean approach:

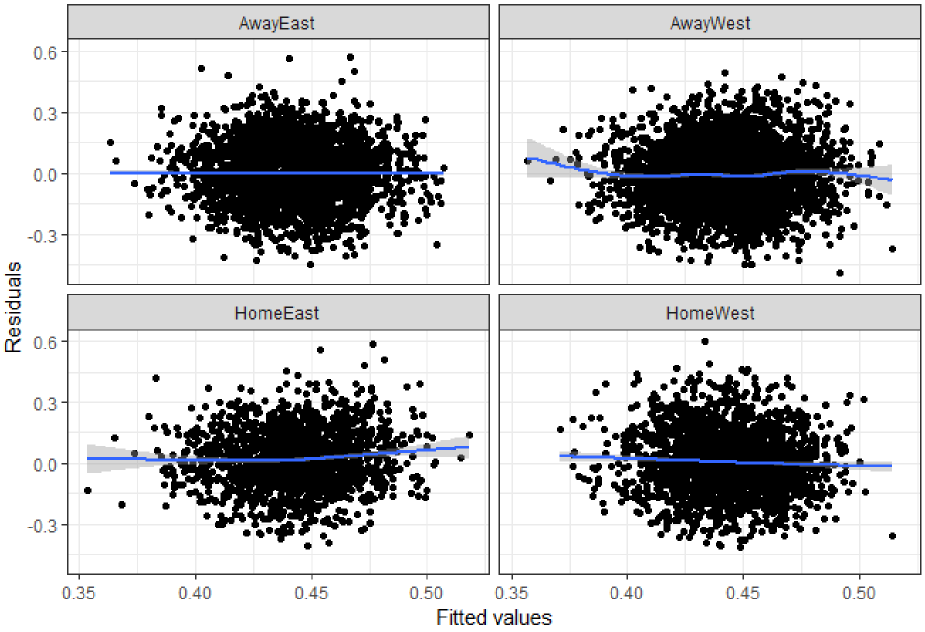


Figure 18. Homogeneity of Errors Boolean Model by jet lag measurement.

b. *Normality of errors:*

The figure below assess the Normality of errors in the linear model from OPA. This histogram displays the frequencies distribution of the error residuals, it also shows in black color a bell curve representing a normal distribution:

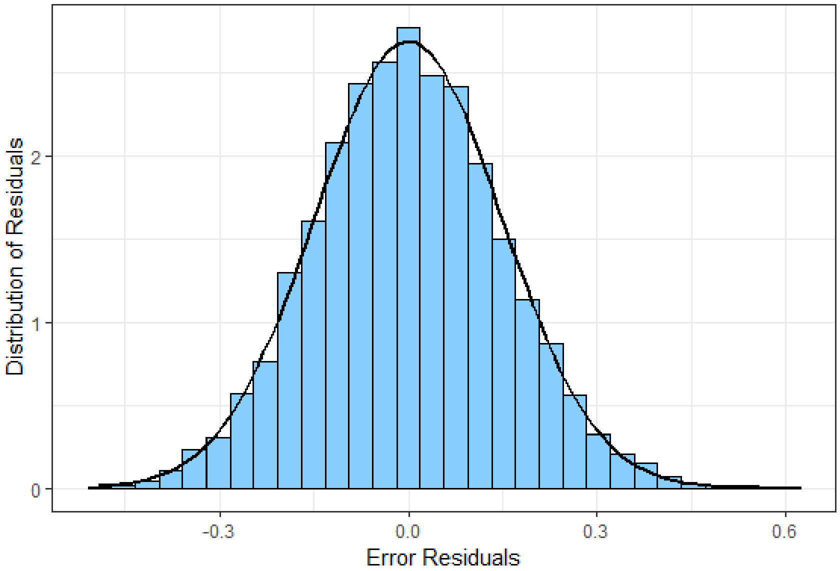


Figure 19. Distribution of error residuals Boolean Model

The figure below complements the assessment of normality of errors by doing it errors across the different levels of measurement for the jet lag presence in the Boolean linear model from OPA. It shows the frequencies distribution of errors across the fitted values for the linear model in each one of the main categories under evaluation for the Boolean approach:

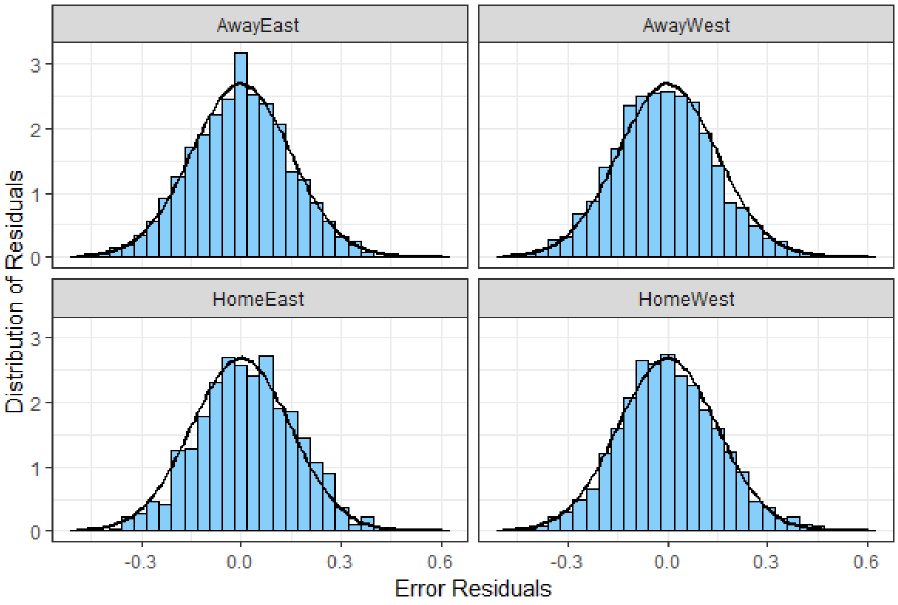


Figure 20. Distribution of error residuals Boolean Model by jet lag category.

The previous visual analysis is complemented with a Quantile-Quantile probability plot where is possible to compare a theoretical normal distribution with the distribution of OPA errors under evaluation:

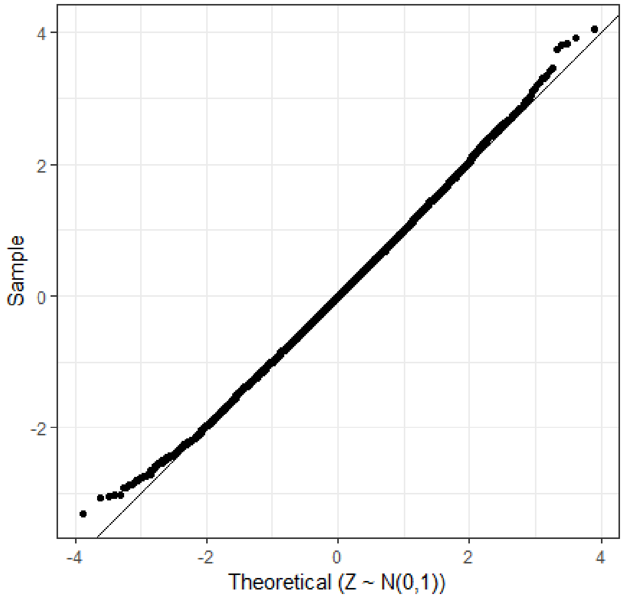


Figure 21. Quantile-Quantile plot to evaluate normality boolean model.

As in the previous analysis, the figure below complements the assessment of normality of errors across the different levels of measurement for the jet lag presence in the Boolean linear model from OPA:

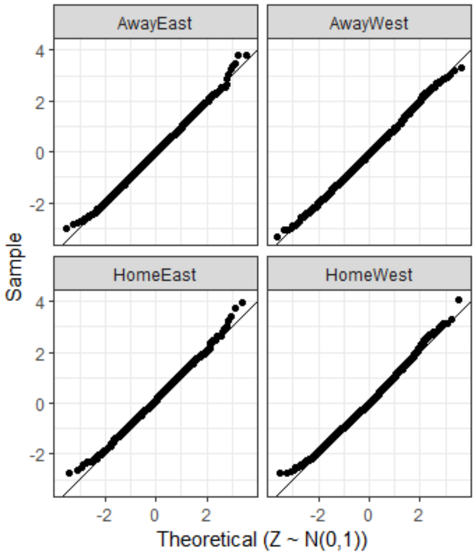


Figure 22. Quantile-Quantile plot to evaluate normality boolean model by category of jet lag.

Findings: From the previous artefacts extracted as results from this analysis is possible to evidence that the multivariate linear regression model created to evaluate the Boolean approach of jet lag effects over offensive performance complies with the definition of the goodness of fit (MacGillivray, Utts, & Heckard, 2013). It means that the assumptions underlying the linear regression model such as the homogeneity distribution of errors and the Normality of errors were met.

2. *Categorical Linear Model*

a. *Homogeneity of errors:*

The figure below assess the Homogeneity of errors in the linear model from OPA. It shows the distribution of errors across the fitted values for the linear model**:**

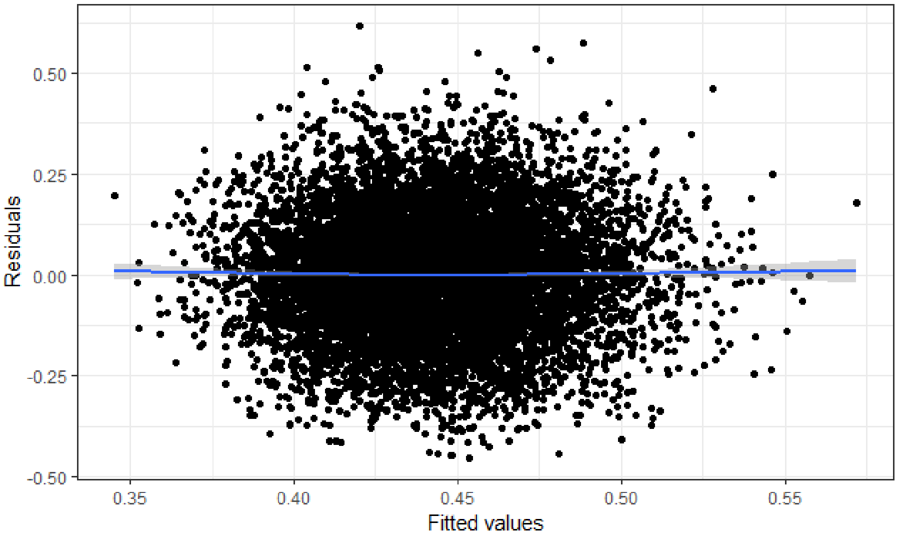


Figure 23. Homogeneity of Errors Categorical Model

The figure below assess the Homogeneity of errors across the different levels of measurement for the jet lag presence in the categorical linear model from OPA. It shows the distribution of errors across the fitted values for the linear model in each one of the main categories under evaluation for the Boolean approach:

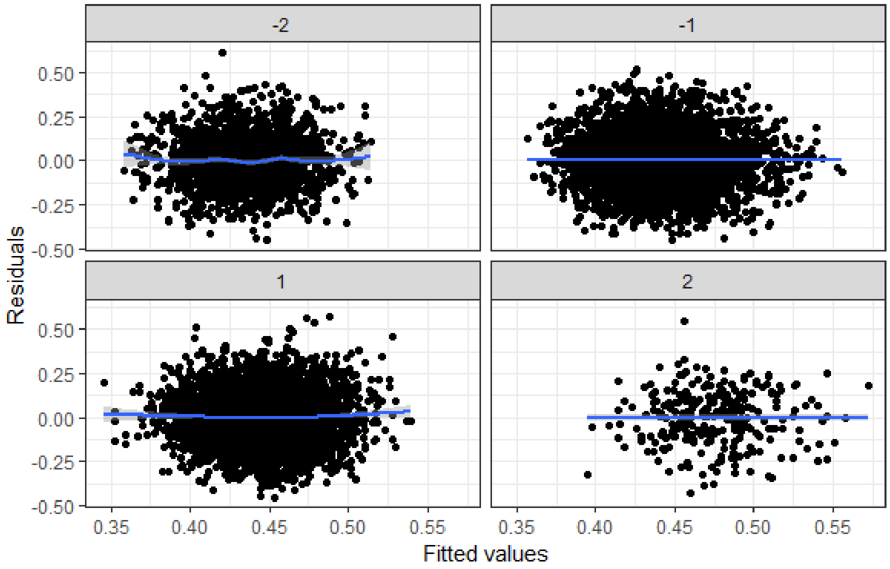


Figure 24. Homogeneity of Errors Boolean Model by jet lag levels.

b. *Normality of errors:*

The figure below assess the normality of errors in the linear model from OPA. This histogram displays the frequencies distribution of the error residuals, it also shows in black color a bell curve representing a normal distribution

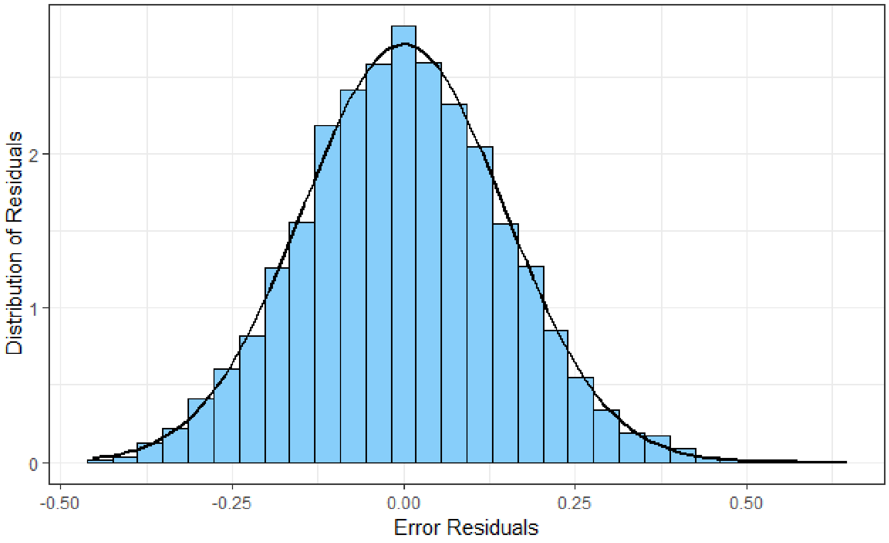


Figure 25. Distribution of error residuals Categorical Model

The figure below complements the assessment of normality of errors by doing it across the different levels of measurement for the jet lag presence in the Categorical linear model from OPA. It shows the frequencies distribution of errors across the fitted values for the linear model in each one of the main categories under evaluation for the Categorical approach:

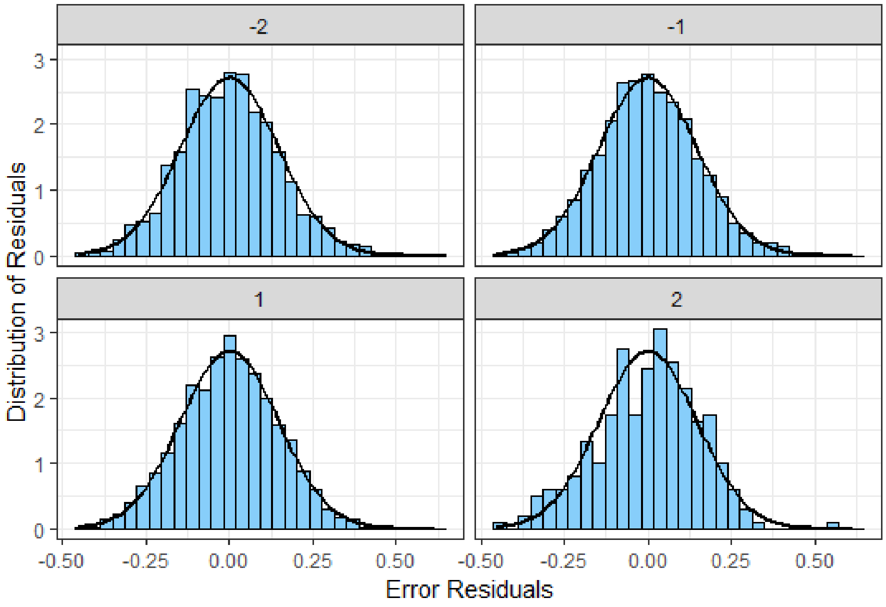


Figure 26. Distribution of error residuals Categorical Model by jet lag level.

The previous visual analysis is complemented with a Quantile-Quantile probability plot where is possible to compare a theoretical normal distribution with the distribution of OPA errors under evaluation:

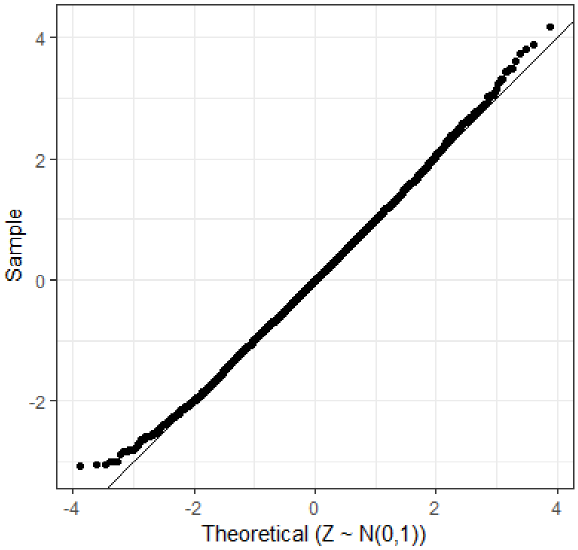


Figure 27. Quantile-Quantile plot to evaluate normality Categorical model.

As in the previous analysis, the figure below complements the assessment of normality of errors across the different levels of measurement for the jet lag presence in the Categorical linear model from OPA:

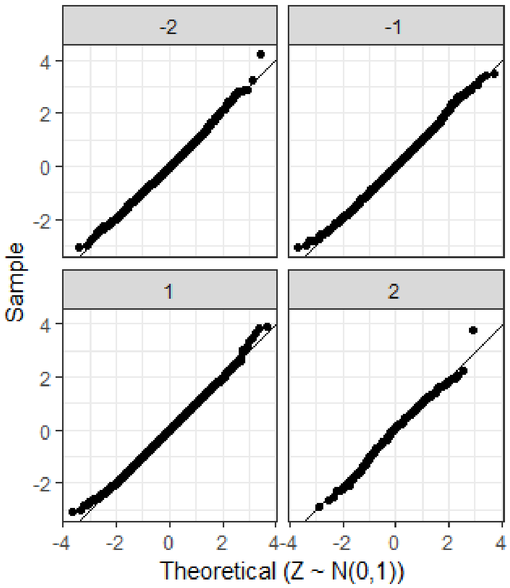
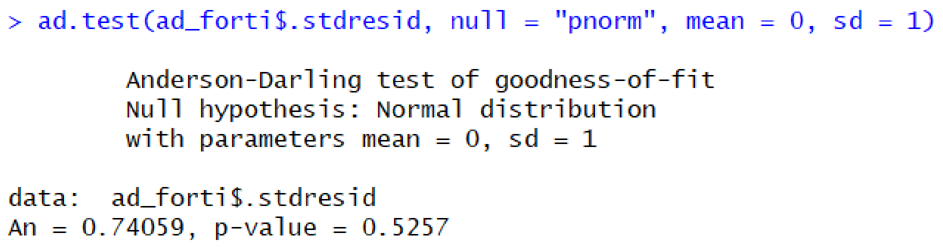


Figure 28. Quantile-Quantile plot to evaluate normality Categorical model by jet lag level.

From the previous images is possible to observe some special behavior over the data accounted for the category +2 which represent 2 hours of jet lag earthwards. This behavior is evaluated by using a normality test known as Anderson darling test which was also described during the methodology section. This test will allow evaluating whether or not the distribution of errors from OPA versus this category follow a normal distribution.



```
> ad.test(ad_forti$.stdresid, null = "pnorm", mean = 0, sd = 1)

        Anderson-Darling test of goodness-of-fit
        Null hypothesis: Normal distribution
        with parameters mean = 0, sd = 1

data:  ad_forti$.stdresid
An = 0.74059, p-value = 0.5257
```

The value gathered from the Anderson darling test evidence that the data is normal distributed due to there is not enough evidence to reject the null hypothesis which states that the data for the category +2 of jetlag is normal distributed.

Findings: From the previous artefacts extracted as results from this analysis is possible to evidence that the multivariate linear regression model created to evaluate the Categorical approach of jet lag effects over offensive performance complies with the definition of goodness of fit (MacGillivray, Utts, & Heckard, 2013). It means that the assumptions underlying the linear regression model such as the homogeneity distribution of errors and the Normality of errors were met.

4. Discussion

From the results previously shown, was possible to evidence that the data the project is working with is balanced, moreover, the explanatory and outcome variables under evaluation was shown to be normally distributed. This is not surprising due to the nature of the variables and the way as they were calculated during the transformation steps. Moreover, the linear regression models created were evaluated in terms of their assumptions. From there was possible to determine that the information provided by the linear models is useful and trustful due to it meets the assumptions of homogeneity of errors and Normality of errors. Therefore, the results presented by the analysis of each offensive statistic can be considered as statistically accurate. These previous results are not surprising, due to the present analysis correspond and extension of previous work (Song, Severini, & Allada, 2017) executed under the same source data (up to 20 years) by using similar linear regression models, which provides a sense of truthfulness.

The following step of the analysis was executed over each offensive statistic.

- The first statistic to be evaluated was the Batting Average (hits/at-bats). In terms of Boolean approach, only the Net_Elo parameter reached statistical significance $\boldsymbol{NetElo}\,(0.0189\%, P = \sim 0.0)$. On the one hand, the Home effect did not reach any statistical significance neither westward $(-0.077\%, P = 0.8010)$ nor eastward $(0.024\%, P = 0.5501)$. On the other hand, the Away effect showed similar results where neither eastward effect $(-0.00695\%, P = 0.9887)$ nor westward $(0.57\%, P = 0.1366)$ reached any statistical significance. However, slightly different results were gathered when the analysis was executed by applying the categorical approach. Although the Net_Elo remained significant $\boldsymbol{NetElo}$ $(0.0189\%, P = \sim 0.0)$ with a similar magnitude of relationship, the Away effect showed a variation in the westward component with the category -1 ($\boldsymbol{AwayWest_{-1}}(0.90\%, P = 0.023)$) which reached significance. Moreover, the other categories either in Away or Home did not reach significance.

- The second statistic to be evaluated was the On Base Performance (calculated as Hits + Walks + Hit by pitch divided by at-bats + walks + hit by pitch + sacrifices). In terms of Boolean approach, only the Net_Elo parameter reached statistical significance $\boldsymbol{NetElo}\,(0.028\%, P = \sim 0.0)$. On the one hand, the Home effect did not reach any statistical significance neither westward $(-0.11\%, P = 0.7038)$ nor eastward $(0.7411\%, P = 0.0575)$. On the other hand, the Away effect showed different results where the westward $\boldsymbol{AwayWest}\,(0.731\%, P = 0.0496)$ effect reached significance and eastward $(0.29\%, P = 0.5394)$ effect did not. However, surprisingly, completely different results were gathered when the analysis was executed by applying the categorical approach. Although the Net_Elo remained significant $\boldsymbol{NetElo}$ $(0.028\%, P = \sim 0.0)$ with a similar magnitude of relationship, the Away effect Eastwards with a level of 2 $\boldsymbol{AwayEast_2}\,(2.5\%, P = 0.03)$ reached significance. Moreover, all the categories belonging to Home Eastwards $\boldsymbol{HomeEast_1}(1.6\%, P = 0.002)$, $\boldsymbol{HomeEast_2}(2.8\%, P = 0.003)$ and Home Westwards category with level -1 $\boldsymbol{HomeWest_{-1}}$ $(-0.669\%, P = 0.04)$ reached significance. The other categories either in Away or Home did not reach significance.

- The third statistic to be evaluated was the Slugging Percentage (calculated as Singles + 2 Doubles + 3Triples + 4Home Runs)/at-bats). In terms of Boolean approach, only the Net_Elo parameter reached statistical significance $\boldsymbol{NetElo}$ $(0.0434\%, P = \sim 0.0)$. On the one hand, the Home effect did not reach any statistical significance neither westward

$(0.460\%, P = 0.3768)$ nor eastward $(0.293\%, P = 0.6680)$. On the other hand, the Away effect showed similar results where nor the eastward effect $(-0.0612\%, P = 0.9416)$ and nor westward $(1.23\%, P = 0.0593)$ reached any statistical significance. Moreover, as was expected, the same results were gathered when the analysis was executed by applying the categorical approach. The Net_Elo remained significant $\boldsymbol{NetElo}$ $(0.0434\%, P = \sim 0.0)$ with a similar magnitude of relationship, and the other categories either in Away or Home did not reach any significance.

- The fourth statistic to be evaluated was the Offensive Average where in terms of the boolean approach, only the Net_Elo parameter reached statistical significance $\boldsymbol{NetElo}$ $(0.050\%, P = \sim 0.0)$. On the one hand, the Home effect did not reach any statistical significance neither westward $(0.458\%, P = 0.3506)$ nor eastward $(0.98\%, P = 0.1284)$. On the other hand, the Away effect showed different results where the westward $\boldsymbol{AwayWest}$ $(1.4\%, P = 0.0229)$ effect reached significance and the eastward $(0.309\%, P = 0.6951)$ effect did not. However, completely different results were gathered when the analysis was executed by applying the categorical approach. Although the Net_Elo remained significant $\boldsymbol{NetElo}$ $(0.0498\%, P = \sim 0.0)$ with a similar magnitude of relationship, the Away effect Eastwards with a level of 2 $\boldsymbol{AwayEast_2}$ $(2.85\%, P = 0.047)$ reached significance. Moreover, all the categories belonging to Home Eastwards $\boldsymbol{HomeEast_1}$ $(2.09\%, P = 0.001)$, $\boldsymbol{HomeEast_2}$ $(4.7\%, P = 0.003)$ and the Home Westwards category with level -1 $\boldsymbol{HomeWest_{-1}}$ $(-1.3\%, P = 0.016)$reached significance. The other categories either in Away or Home did not reach significance.

- The fifth statistic was the Offensive Performance Average (OPA) where in terms of Boolean approach, only the Net_Elo parameter reached statistical significance $\boldsymbol{NetElo}$ $(0.043\%, P = \sim 0.0)$. On the one hand, the Home effect did not reach any statistical significance neither westward $(0.334\%, P = 0.4488)$ nor eastward $(0.692\%, P = 0.2323)$. On the other hand, the Away effect showed different results where eastward effect $(0.154\%, P = 0.8276)$ did not reach significance, but westward $\boldsymbol{AwayWest}$ $(1.17\%, P = 0.0333)$ effect reached statistical significance. However, slightly different results were gathered when the analysis was executed by applying the categorical approach. Although the Net_Elo remained significant $\boldsymbol{NetElo}$ $(0.043\%, P = \sim 0.0)$ with a similar magnitude of relationship, the Home effect showed a variation in the eastward component where both categories $\boldsymbol{HomeEast_1}$ $(1.54\%, P = 0.008)$, $\boldsymbol{HomeEast_2}$ $(3.709\%, P = 0.010)$ reached statistical significance. Moreover, the other categories either in Away or Home did not reach significance.

- Finally, the last variable under evaluation was Runs Scored. In terms of Boolean approach, only the Net_Elo parameter reached statistical significance $\boldsymbol{NetElo}$ $(0.88\%, P = \sim 0.0)$. On the one hand, the Home effect did not reach any statistical significance neither westward $(3.24\%, P = 0.7226)$ nor eastward $(10.96\%, P = 0.3609)$. On the other hand, the Away effect showed different results where the westward $\boldsymbol{AwayWest}$ $(27.27\%, P = 0.0173)$ effect reached significance and the eastward $(-2.26\%, P = 0.0173)$ effect did not. However, slightly different results were gathered when the analysis was executed by applying the categorical approach. The Net_Elo remained significant $\boldsymbol{NetElo}$ $(0.87\%, P = \sim 0.0)$ with a similar magnitude of relationship, and the other categories either in Away or Home did not reach any significance.

Comparing these results to prior work, our categorical approach seems to produce similar results as the observed in the research of Song (2017), where according to the previous results the effects detected on home team were more robust than on away team, also, the findings on this prior studies shown that most major jetlag effects were evident in eastwards travels. These conclusions are aligned to the information provided by the categorical model developed in this project, where 8 from 11 significant parameters were detected eastwards and also the same ratio applies to significant effects of Home over Away effects. Moreover, as was mentioned in the prior work, some isolated effects of westward travel were observed. This same behaviour was observed in the categorical model, were for the Batting Average(BA), On Base Performance(OBP) and Offensive Average(OA) the west effect of level -1 was found as significant. Finally, the inclusion of different levels of jet lag allows us to evidence that the coefficients among levels increase as the level goes up in a similar ratio among statistics. This effect was observed in Offensive Average where the level 1 accounted for a coefficient of 0.0209035757 while the level 2 accounted 0.0474526845. Offensive Performance Average where level 1 accounted for 0.0154065709 and level 2 0.0370928965. Offensive batting performance where level 1 accounted for 0.0166776917 and level 2 0.0286894053.

However, the present work has shown some clear limitations regarding the Boolean model implemented which differs completely from the results observed in prior works. While the prior work shown the jet lag effects experienced after eastwards travel and for Home teams as more significant, the Boolean model developed in the present project resulted in Away teams and westward travels as more significant. It can be associated with the new inclusion of the difference of relative skills among the teams, which can be affected by the results obtained by adding more significance to the effect of Away and Westward jetlag. On the other hand, this difference could be related to variations in the interpretation of assumptions coming from the previous work which make the model developed weak and not accurate. For instance, the assumption of 1-hour compensation by every 24 hours had different interpretations when it was applied, which could easily lead to variations in the results and errors in the model implemented.

Some key recommendations based on the experience gathered from the present project to keep under evaluation would be focused on the improvement of assumptions interpretation coming from prior work. Several assumptions should be defined when a data analytics project is executed, some of them were followed as prior efforts executed them, however not all the assumptions were clearly defined in the research papers collected. Therefore, to pay special attention from the begin of the project over assumptions such as the way of handling outliers values of jet lag, the way to calculate the 24h of compensation applied over the jetlag and also the way to include the categorical variables into the multivariate linear regression model can be really significant over the accuracy and reliability of the model produced, leading to better results. On the other hand, the next steps that can be followed to go further can be to evaluate the jet lag effects at a player level by evaluating the performance of players through a season when they are affected by travelling across time zones. Moreover, the comparison between different linear regression models can leverage some additional insights regarding the jet lag relationships.

5. Conclusions

The previous effort took into consideration several meaningful variables surrounding the jet lag impact over offensive performance such as the direction of travel, the effect of home or away advantage and the relative skill level of each team. It also based on previous efforts in the same area by extending the amount of data evaluated and considering additional meaningful statistics recommended by other studies. Finally, two different approaches both boolean and categorical were applied to gather additional awareness regarding the likely relationship between variables. Following the previous structure, the conclusions of the present effort may be drawn from the perspective of each one of the models followed.

In terms of the multivariate linear regression model using the jet lag boolean representation, significant effects of travelling across several times zones were found when teams travelled westwards and played as away. Similar effects were observed in four over six of the offensive statistics under evaluation. These results differ from previous research as shown during the environmental scan where eastward flights and home teams were found more significant. These results can be attributed to the effect of net Elo, any variation during the calculation process executed over the data or any misinterpretation of assumptions made in the research followed.

In contrast, the multivariate linear regression model using the jet lag categorical representation shown different results. This model showed more significant effects after travelling eastward than westward. Moreover, it displayed a predominance for home playing over away. These results seem to be aligned to previous results gathered not only from different experiments in medical assessments but also in the research-based to develop the present project. On the other hand, as was mentioned in the discussion, the inclusion of different levels of jet lag allows us to evidence a linear relationship between levels of jet lag where the effects seem to double when the effect is increased in one additional level.

Unsurprisingly, both models agreed in the significant effect inserted by the net difference between the relative skills among teams. This variable was shown as significant in all the offensive statistics across both experiments. This result could be explained due to the nature of performance accounted by this measurement, where positive difference implies better teams accounting better statistics and negative net Elo values implies teams with the worst statistics.

The significance of this work can be explained in terms of the decision to be made by coaches and team owners when need to face games after crossing several time zones. For instance, a team owner may request to schedule the travel of some specific player before the whole team in order to fit the new time zone for instances when games are played eastward. On the other hand, coaches may take a different decision regarding the behaviour of specific players during the training previous to games to be played at Home in order to avoid strong modifications of his circadian rhythms which may affect the results. Finally, these finding can be considered by the board of leagues from different sports who may change the patterns of how the sequence of games is decided in order to ensure better performance of the teams by avoiding the influence of extenuated travels. These decisions from the board may increase the competitiveness of the league leading to most interesting and attractive games which at the end will benefit the league and the teams.

Potential further work can be developed by evaluating deeply the possible linear dependency between the levels of jet lag, by considering domains where more time zones are crossed and comparing the results among levels. Moreover, information about each travel done by each team should be included to consider transferences among aeroplanes, waiting time in airports and other issues faced during each movement. Finally, further analysis over individual players is strongly recommended because as has been stated before, jet lag effects are considerably different among people where neither the same effects and nor the same magnitude of effects is experienced equally by different human beings.

6. REFERENCES

Baumer, B., & Zimbalist., A. (2013). The sabermetric revolution: Assessing the growth of analytics in baseball. *University of Pennsylvania Press*. Philadelphia.

Bennett, J., & Flueck, J. (1983). An Evaluation of Major League Baseball Offensive Performance Models . . *The American Statistician, 37, 1*, 76–82. doi:https://doi.org/10.1080/00031305.1983.10483093

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Fowler, M. K., Crowcroft, E., Mendham, E., Miller, E., Sargent, E., & ... Duffield, E. (2017). Greater Effect of East versus West Travel on Jet Lag, Sleep, and Team Sport Performance. . *Medicine & Science in Sports & Exercise, 49, 12*, 2548–2561. doi:https://doi.org/10.1249/MSS.0000000000001374

Glickman, M. E., & Jones, A. C. (1999). (Rating the chess rating system. *Chance, 12, 2*, 21–28.

Graeber, R., Kryger, M. H., Roth, T., & Dement, W. (1994). Jet lag and sleep disruption. *Principles and practice of sleep medicine.*, 463-70. Philidelphia: WB Saunders.

Klein, K., Wegmann, H., & Athanassenas, G. (1976). Air operations and circadian performance rhythms. *Aviat Space Environ Med, 47*, 221-30.

Lehmann, R., & Wohlrabe, K. (2017). Who is the “Journal Grand Master”? A new ranking based on the Elo rating system. *Journal of Informetrics, 11, 3*, 800–809. doi:https://doi.org/10.1016/j.joi.2017.05.004

Liu, S., Tian, X., & Zhang, Z. (2010). Process planning knowledge discovery in the process database. *In 2010 International Conference on Computer Application and System Modeling (ICCASM 2010). Vol. 11*, pp. V11–370–V11–373. IEEE. doi:https://doi.org/10.1109/ICCASM.2010.5623186

MacGillivray, H., Utts, J., & Heckard, R. (2013). Mind on Statistics. *2nd Australian and New Zealand. Cengage Learning*. Australia.

Mcguckin, T., Sinclair, W., Sealey, R., & Bowman, P. (2014). The effects of air travel on performance measures of elite Australian rugby league players. . *European Journal of Sport Science, 14* (S1), 116–122. doi:https://doi.org/10.1080/17461391.2011.654270

Pankin, M. D. (1978). Evaluating Offensive Performance in Baseball. *Operations Research, 26, 4*, 610–619. doi:https://doi.org/10.1287/opre.26.4.610

Recht, L. D., Lew, R. A., & Schwartz, W. J. (1995). Baseball teams beaten by jet lag. *Nature, 377, 6550*, 583–583. doi:https://doi.org/10.1038/377583a0

Redfern, P., Minors, D., & Waterhouse, J. (1994). Circadian rhythms, . Chronobiollnt . *jet lag, and chronobiotics: an overview, 11*, 253- 65.

Retrosheet, O. (2018, 11 21). *Retrosheet*. Retrieved from https://www.retrosheet.org/game.htm

Saunders, G. (2016, March 17). *University of St Andrews.* Retrieved from Digital communications team blog: http://digitalcommunications.wp.st-andrews.ac.uk/2016/03/17/dsdm-agile-project-management-cheat-sheet/

Schwaber, K., & Sutherland, J. (2017). *Scrumalliance.* Retrieved from The Scrum Guide: https://www.scrumalliance.org/learn-about-scrum/the-scrum-guide

Smith, R., Guilleminault, C., & Efron, B. (1997). Circadian rhythms and enhanced athletic performance in the national football league. *Sleep, 20, 5*, 362–365.

Song, A., Severini, T., & Allada, R. (2017). How jet lag impairs Major League Baseball performance. *Proceedings of the National Academy of Sciences of the United States of America, 114, 6*, 1407–1412. doi:https://doi.org/10.1073/pnas.1608847114

Takahashi, T., Sasaki, M., Itoh, H., Sano, H., Yamadera, W., Ozone, M., & Matsunaga, N. (1999). Re-entrainment of circadian rhythm of plasma melatonin on an 8-h eastward flight. , . *Psychiatry and Clinical Neurosciences, 53, 2*, 257–260. . doi:https://doi.org/10.104

Waterhouse, J., Reilly, T., Atkinson, G., & Edwards, B. (2007). Jet lag: trends and coping strategies. *The Lancet, 369, 9567*, 1117–1129. doi:https://doi.org/10.1016/S0140-6736(07)60529-7

Wever, R. (1966). The duration of re-entrainment of circadian rhythms after phase shifts of the Zeitgeber A theoretical investigation. *Journal of Theoretical Biology., 13(1), C*, 187–201. doi:https://doi.org/10.1016/0022-5193(66)90016-6

Youngstedt, S. D., & O’Connor, P. J. (1999). The influence of air travel on athletic performance. *Sports Medicine, 28(3)*, 197-207.

Youngstedt, S., & O’Connor, P. (1999). The Influence of Air Travel on Athletic Performance. . *Sports Medicine, 28, 3*, 197–207. . doi:https://doi.org/10.2165/00007256-199928030-00004

APPENDIX A:

The present project followed an agile methodology to achieve the expected results. As was planned from the begin of the project, 7 sprints were executed through the implementation stage. Each sprint was compound by a sprint meeting with the project supervisor where the work to be done during the sprint was defined. Each meeting accounts for a weekly meeting log that can be observed in appendix A. Moreover, while the project progressed from sprint #1 to sprint #7, the final deliverables were built progressively, and the changes found regarding assumptions not considered from the begin or misinterpretations done at the begin were solved iteratively.

During the execution of the present project, several activities were executed to reach the final objective proposed from the begin of the project. In my personal opinion, the activities that were done better than the average were compound by the coding section developed in python. It was a section that challenged my professional skills in this programing language, therefore I studied and improved significantly my skills to achieve a remarkable level and produce a good result. On the other hand, the toughest group of activities were related to understanding the assumptions made by the previous research followed during this work. It could be influenced due to the lack of domain knowledge from the people involved in the project or the lack of clear information provided by the authors. For these reasons, I consider the project could be affected by misinterpretations or biases regarding the way as the assumptions were implemented through the project.

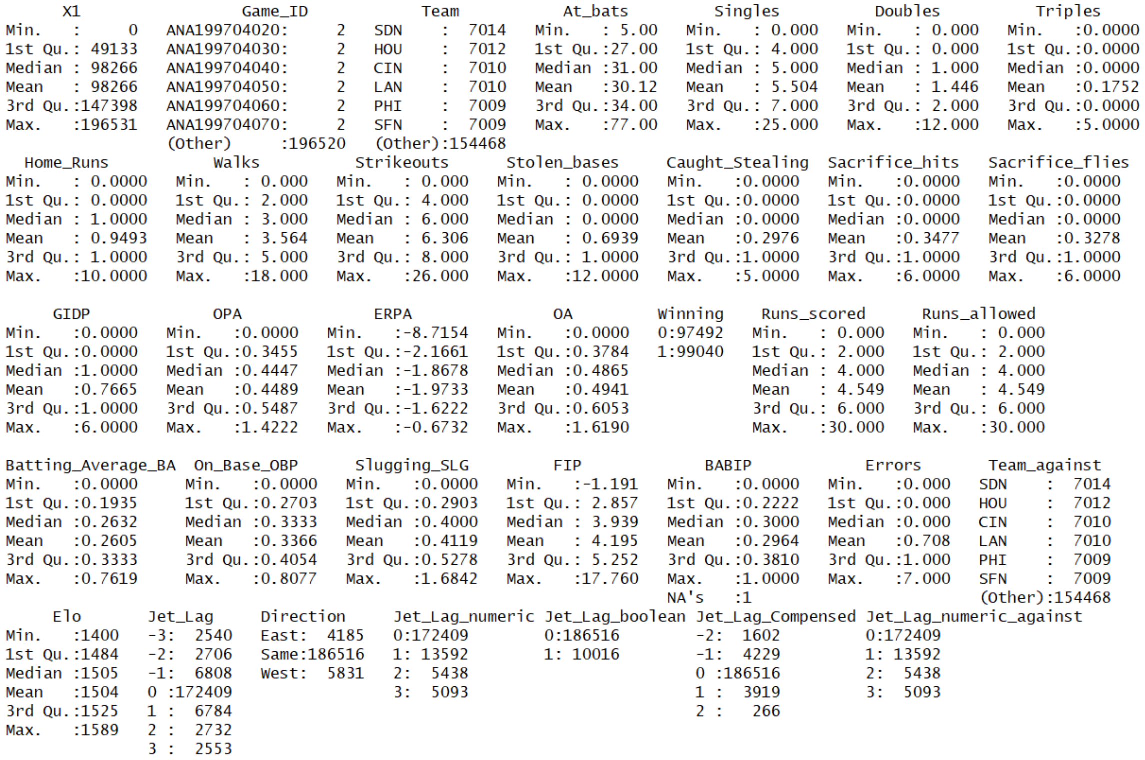
Regarding problems encountered during the project, I can point out that the project faced issues when the jet lag variable was inserted. Although the number of games considered under jet lag effects matched perfectly with the numbers provided by the previous research, some values of jetlag differed among our work and the previous research. It could be produced by any assumption misinterpreted or any error during the applying of the jet lag equation over the dataset. To handle this issue, several possible interpretations of the assumption of jet lag compensation were applied including the consideration of exactly 24 hours and the consideration of day subtraction.

The hardest part of this project was to come up with a feasible approach to extract and transform the data provided by the Retrosheet organization. To sort this section, additional development on my python skills were required. However, additional further development is required in python programming due to the extension and variety of possible approaches which can be implemented. Similarly, further knowledge of statistical data analysis would be really valuable due to the importance of this area under the analysis executed. On the other hand, the most important thing learnt from this project was the concepts behind the multivariate linear regression analysis and the ways to validate whether or not the model created to meet the assumptions to be considered as valid. This knowledge represented something new for my professional development. In terms of project management and research, the most valuable knowledge gathered from this project was the importance of the iterative approach used to produce the results by breaking down the development and saving time during the phases of the project.

In terms of strengths, I consider the present project came up with a feasible and useful solution to extract and process the online data available in Retrosheet organization. This approach can be used for the following projects in order to set a starting point for further developments or analysis. On the other hand, the weaknesses of this project can be reflected in several sections. The first weakness may be represented by the differences among the values of jet lag inserted by the project and the results provided by the previous research. As was mentioned before, the next weakness is related to the assumption interpretations done during the project. The previous research implemented some particular considerations which may change the final results if they were not correctly implemented. Finally, during the analysis stage, the project did not include the information regarding games without any effect of jet lag. It could be considered a biased decision, but it was done by following the same methodology from the previous research.

Regarding next steps based on the learnings provided by the project, thoroughly analysis over the possible linearity stated among levels of jet lag could be beneficial to understand the type of relationship existent among them. Moreover, the importance of considering each player as independent would be really important due to the differences in jet lag effects among people which can lead to new valuable awareness.

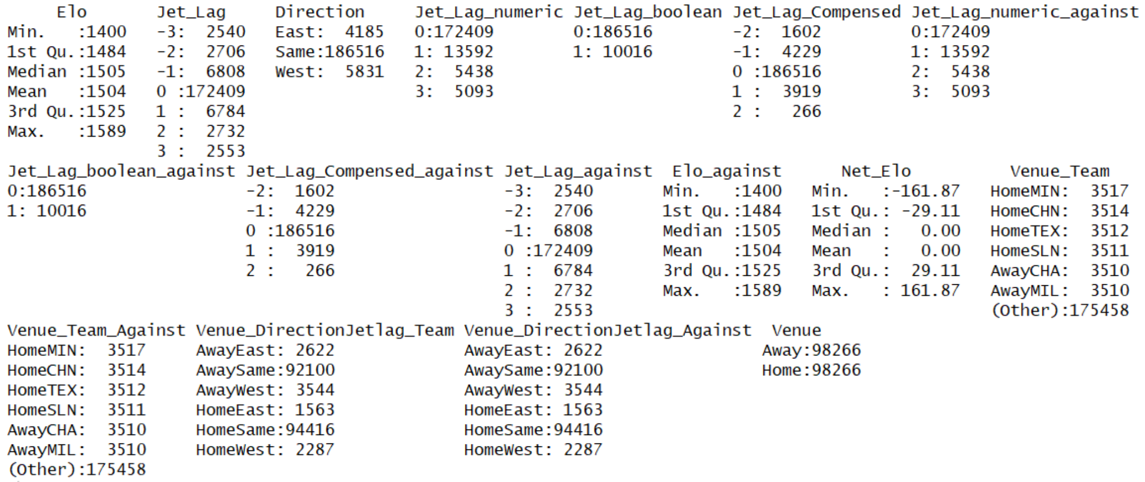
APPENDIX B:



```
      X1                   Game_ID                Team            At_bats         Singles          Doubles          Triples      
Min.   :     0   ANA199704020:     2   SDN    :  7014   Min.   : 5.00   Min.   : 0.000   Min.   : 0.000   Min.   :0.0000  
1st Qu.: 49133   ANA199704030:     2   HOU    :  7012   1st Qu.:27.00   1st Qu.: 4.000   1st Qu.: 0.000   1st Qu.:0.0000  
Median : 98266   ANA199704040:     2   CIN    :  7010   Median :31.00   Median : 5.000   Median : 1.000   Median :0.0000  
Mean   : 98266   ANA199704050:     2   LAN    :  7010   Mean   :30.12   Mean   : 5.504   Mean   : 1.446   Mean   :0.1752  
3rd Qu.:147398   ANA199704060:     2   PHI    :  7009   3rd Qu.:34.00   3rd Qu.: 7.000   3rd Qu.: 2.000   3rd Qu.:0.0000  
Max.   :196531   ANA199704070:     2   SFN    :  7009   Max.   :77.00   Max.   :25.000   Max.   :12.000   Max.   :5.0000  
                 (Other)     :196520   (Other):154468                                                                     
  Home_Runs           Walks            Strikeouts        Stolen_bases      Caught_Stealing   Sacrifice_hits    Sacrifice_flies 
Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
1st Qu.: 0.0000   1st Qu.: 2.000   1st Qu.: 4.000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
Median : 1.0000   Median : 3.000   Median : 6.000   Median : 0.0000   Median :0.0000   Median :0.0000   Median :0.0000  
Mean   : 0.9493   Mean   : 3.564   Mean   : 6.306   Mean   : 0.6939   Mean   :0.2976   Mean   :0.3477   Mean   :0.3278  
3rd Qu.: 1.0000   3rd Qu.: 5.000   3rd Qu.: 8.000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
Max.   :10.0000   Max.   :18.000   Max.   :26.000   Max.   :12.0000   Max.   :5.0000   Max.   :6.0000   Max.   :6.0000  

     GIDP              OPA               ERPA                OA             Winning      Runs_scored       Runs_allowed   
Min.   :0.0000   Min.   :0.0000   Min.   :-8.7154   Min.   :0.0000   0:97492   Min.   : 0.000   Min.   : 0.000  
1st Qu.:0.0000   1st Qu.:0.3455   1st Qu.:-2.1661   1st Qu.:0.3784   1:99040   1st Qu.: 2.000   1st Qu.: 2.000  
Median :1.0000   Median :0.4447   Median :-1.8678   Median :0.4865             Median : 4.000   Median : 4.000  
Mean   :0.7665   Mean   :0.4489   Mean   :-1.9733   Mean   :0.4941             Mean   : 4.549   Mean   : 4.549  
3rd Qu.:1.0000   3rd Qu.:0.5487   3rd Qu.:-1.6222   3rd Qu.:0.6053             3rd Qu.: 6.000   3rd Qu.: 6.000  
Max.   :6.0000   Max.   :1.4222   Max.   :-0.6732   Max.   :1.6190             Max.   :30.000   Max.   :30.000  

Batting_Average_BA  On_Base_OBP       Slugging_SLG          FIP               BABIP              Errors         Team_against   
Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :-1.191   Min.   :0.0000   Min.   :0.000   SDN    :  7014  
1st Qu.:0.1935     1st Qu.:0.2703   1st Qu.:0.2903   1st Qu.: 2.857   1st Qu.:0.2222   1st Qu.:0.000   HOU    :  7012  
Median :0.2632     Median :0.3333   Median :0.4000   Median : 3.939   Median :0.3000   Median :0.000   CIN    :  7010  
Mean   :0.2605     Mean   :0.3366   Mean   :0.4119   Mean   : 4.195   Mean   :0.2964   Mean   :0.708   LAN    :  7010  
3rd Qu.:0.3333     3rd Qu.:0.4054   3rd Qu.:0.5278   3rd Qu.: 5.252   3rd Qu.:0.3810   3rd Qu.:1.000   PHI    :  7009  
Max.   :0.7619     Max.   :0.8077   Max.   :1.6842   Max.   :17.760   Max.   :1.0000   Max.   :7.000   SFN    :  7009  
                                                                      NA's   :1                         (Other):154468  
     Elo         Jet_Lag       Direction      Jet_Lag_numeric Jet_Lag_boolean Jet_Lag_Compensed Jet_Lag_numeric_against
Min.   :1400   -3:  2540   East:  4185   0:172409        0:186516        -2:  1602        0:172409               
1st Qu.:1484   -2:  2706   Same:186516   1: 13592        1: 10016        -1:  4229        1: 13592               
Median :1505   -1:  6808   West:  5831   2:  5438                        0 :186516        2:  5438               
Mean   :1504   0 :172409                 3:  5093                        1 :  3919        3:  5093               
3rd Qu.:1525   1 :  6784                                                 2 :   266                               
Max.   :1589   2 :  2732                                                                                         
               3 :  2553                                                                                         
```



```
     Elo         Jet_Lag       Direction      Jet_Lag_numeric Jet_Lag_boolean Jet_Lag_Compensed Jet_Lag_numeric_against
Min.   :1400   -3:  2540   East:  4185   0:172409        0:186516        -2:  1602        0:172409               
1st Qu.:1484   -2:  2706   Same:186516   1: 13592        1: 10016        -1:  4229        1: 13592               
Median :1505   -1:  6808   West:  5831   2:  5438                        0 :186516        2:  5438               
Mean   :1504   0 :172409                 3:  5093                        1 :  3919        3:  5093               
3rd Qu.:1525   1 :  6784                                                 2 :   266                               
Max.   :1589   2 :  2732                                                                                         
               3 :  2553                                                                                         
```

```
Jet_Lag_boolean_against Jet_Lag_Compensed_against Jet_Lag_against  Elo_against       Net_Elo            Venue_Team    
0:186516                -2:  1602                 -3:  2540       Min.   :1400   Min.   :-161.87   HomeMIN:  3517  
1: 10016                -1:  4229                 -2:  2706       1st Qu.:1484   1st Qu.: -29.11   HomeCHN:  3514  
                        0 :186516                 -1:  6808       Median :1505   Median :   0.00   HomeTEX:  3512  
                        1 :  3919                 0 :172409       Mean   :1504   Mean   :   0.00   HomeSLN:  3511  
                        2 :   266                 1 :  6784       3rd Qu.:1525   3rd Qu.:  29.11   AwayCHA:  3510  
                                                  2 :  2732       Max.   :1589   Max.   : 161.87   AwayMIL:  3510  
                                                  3 :  2553                                        (Other):175458  
Venue_Team_Against Venue_DirectionJetlag_Team Venue_DirectionJetlag_Against  Venue      
HomeMIN:  3517     AwayEast: 2622             AwayEast: 2622                Away:98266  
HomeCHN:  3514     AwaySame:92100             AwaySame:92100                Home:98266  
HomeTEX:  3512     AwayWest: 3544             AwayWest: 3544                            
HomeSLN:  3511     HomeEast: 1563             HomeEast: 1563                            
AwayCHA:  3510     HomeSame:94416             HomeSame:94416                            
AwayMIL:  3510     HomeWest: 2287             HomeWest: 2287                            
(Other):175458                                                                          
```

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 26/02/2019 **Meeting No: 01**

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).
UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** …………………………

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. This was the first meeting

Items for discussion (noted by student before mandatory supervisory meeting):

1. Defining the scope of the project

2. Understand about the source of data and how it should be downloaded.

3. Identify the relevant literature to be reviewed.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Read all the literature review suggested.

2. Have a high level view and idea about retrosheet.org website and data

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 05/03/2019 **Meeting No: 02**

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).
UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** ……………………………………

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Read all the literature material suggested by the supervisor.

2. Navigated through the retrosheet.org website which contains info about the data.

Items for discussion (noted by student before mandatory supervisory meeting):

1. Several doubts to be clarified regarding the literature results and methodology

2. How jet lag should be defined for the research?

3. What is the approach to be taken when using the data

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Identify the methodology used by the target research paper

2. Identify the additional data and metrics to be used in this research.

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 12/03/2019 **Meeting No:** 03

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).
UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** ……………………………

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Methods used by the target research is not clear

Items for discussion (noted by student before mandatory supervisory meeting):

1. Clarifications of the assumptions and methodologies used by the target research paper.

2. Understand how time zones should be used to define jet lag.

3. Understand the other factors that can affect overall jet lag, such as jet lag recovery.

4. Understand the structure of the Project plan.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Identify ways on how jet lag can be represented to add value to the research

2. Identify ways in how the analysis can be divided between me and the other student going forward.

3. Complete the Project Plan.

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 19/03/2019 **Meeting No:** 04

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).
UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** ……………………………

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Completed the Project Plan

Items for discussion (noted by student before mandatory supervisory meeting):

1. Get feedback for the Project plan.

2. Clarify how exactly the previous research can be re-experimented again.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Reconduct the target research and compare the results.

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 08/04/2019 **Meeting No:** 05

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).

UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** …………………………….

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Project plan has been submitted.

2. Target research has been scrutinized.

Items for discussion (noted by student before mandatory supervisory meeting):

1. Clarification regarding several baseball measures used in the target research.

2. Understand the type of statistical analysis to be done on the final data set.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Complete the technical rule excel sheet to decode the notations in the "Event" field of the source file

2. Start extracting all the files from the source.

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 16/04/2019 **Meeting No:** 06

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).

UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** ……………………….

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Technical rule excel sheet has been completed.

2. All data files for all years have been downloaded and appended.

Items for discussion (noted by student before mandatory supervisory meeting):

1. Clarify doubts regarding the definition of "Plate Appearances" and how to calculate it.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Correct the technical rule for Plate appearance and Base running.

2. Check the accuracy of the derived indicators by checking it manually for year 1998.

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 07/05/2019 **Meeting No:** 07

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).

UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** ………………………….

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Validity of the data derived has been checked.

2. All rules have been properly derived.

Items for discussion (noted by student before mandatory supervisory meeting):

1. Clarify doubts regarding the multivariate regression analysis done by the previous research.

2. Clarify doubts regarding jet lag calculation.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Re do the time zone matrix with accurate time zone data.

2. Convert the data in to the regression format.

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 14/05/2019 **Meeting No:** 08

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).

UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** ……………………………

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. All time zones have been properly updated and jet lag calculated.

2. Data has been restructured to be applied in the regression analysis.

Items for discussion (noted by student before mandatory supervisory meeting):

1. Understand how ELO can be included as an additional variable.

2. Understand what offensive metrics should be used.

3. Understand the structure of the final project presentation.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Prepare a first draft of the final presentation.

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Manuel Villamil **Date:** 20/05/2019 **Meeting No:** 09

Project title: EVALUATION OF JETLAG IMPACT OVER OFFENSIVE PERFORMANCE IN BASEBALL TEAMS FROM MAJOR LEAGUE BASEBALL (MLB).

UNIT: IFN701

□ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Dimitri Perrin **Supervisor's Signature:** …………………………

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Completed the final presentation

Items for discussion (noted by student before mandatory supervisory meeting):

1. Get feedback on the final presentation.

2. Understand how the final report should be drafted.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Prepare a first draft of the final report.

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.