“Run For The Money”

*An Exploration on Horse Racing in Hong Kong*

***COMP 30780 Data Science in Practice***

Andrew McGurk (15741011), Dakota Owen (15200586)

School of Computer Science

University College Dublin

*May 11, 2018*

**Abstract**. The aim of this project is to analyse various factors that dictate the multi-layered race tactics and strategies that can have potentially dictate the outcome of a race in horse racing within Hong Kong during the 2014 - ’17 seasons. By examining data collected from these seasons, we can begin to quantify how strategies are in horse racing and measure other variables that might affect how well horses perform. Within this report we shall state the research strategies, key findings, and conclusions of three separate research questions which shall be stated further into this report. Data collection has been sourced from two locations; Kaggle and HK Racing, a website maintained by the South China Morning Post. Finally, we hope to discuss everything about the project from idea conception to final key findings within this report, any mindsets and barriers we came across throughout the project should also be included.

**Declaration**. We (Andrew McGurk, 15741011 and Dakota Owen, 15200586) declare that this assignment is our own work and that we have correctly acknowledged the work of others. This assignment is in accordance with University and School guidance[[1]](#footnote-0) on good academic conduct in this regard.

# **Introduction**

Horse Racing is one of the oldest sporting past times throughout history, dating back as far as the 648BC Greek Olympic Games, it is a monument to human achievement in their ability to tame and live alongside these animals. In the modern day with our advances in technology horses and humans still cohabit as pets, work horses and sporting race horses. As one of the oldest forms of entertainment, horse racing is a sport consisting of at least two horses, ridden by jockeys over a set distance in competition. The formats of these competitions vary greatly from one region to the next however the core ideal is the same everywhere; run fast, and get to the finish line first. While horses are raced purely for sport, a part of the competition’s fierce interest originates from the gambling economy centred around the sport with an estimated €138 billion euro in worldwide market value.

Holding cultural ties to the sport we decided to pursue the idea in order to gain insight and a better understanding of why horse racing is so revered all around the world. In this report we will attempt to answer three research questions discussing the viability of certain strategies and factors that affect a horses performance throughout the race.

Research Question 1; *“Does being a Frontrunner or Slipstreamer result in a better finishing position?”*

Research Question 2; *“Do jockeys affect the performance of horses they ride?”*

Research Question 3; *“Does starting gate offer any advantage to horses in racing?”*

For these research questions we have gone into as much detail as is possible with the data provided. A full description of limitations and barriers to potential better results will be included within the report. Certain key findings have been quite surprising and others confirmed what we had originally believed. Throughout the project, definitions had to be refined multiple times in order to be considered well rounded, an example of a definition that needed to be changed would be Slipstreamers being every horse that isn’t holding 1st position is not effective when it comes to quantifying the effect of slipstreaming in a race.

The first source of data used for the project is from Kaggle, an online community of data scientists and machine learners who organise and analyse a variety of topics with the sole purpose of making the findings available publicly. The second source of data is from our own scraping methods used on HK Racing, a Hong Kong based horse racing and data website owned and maintained by the South China Morning Post.

The remainder of this report is organized as follows. In the next section we shall discuss the motivations and objectives, followed by data wrangling, analysis and results, discussion, conclusions and bibliography.

# **Motivations & Objectives**

## **Background & Motivations**

When first tasked with finding an idea for our Data Science in Practice project, we weren’t entirely sure what we wanted to focus on. One of our first ideas was to perform an in depth exploration and meta-analysis of the work that has been done with regards to the efficient-market hypothesis (EMH). This hypothesis basically states that the prices of securities in the market represent all available information on those securities. This implies that it is impossible to outperform the market after adjusting for the risk of your investments. While very intriguing, we quickly realized it would be incredibly difficult to perform anything more than a surface level analysis of a market theory that has existed for half a century with dozens of associated publications.

With our original project scrapped, it was back to the drawing board. After some discussion, we discovered a mutual interest in horse racing since both project members grew up in areas famous for their races. A few brief searches later and we discovered that not only would it provide us with some intriguing research, but the publicly available data was surprisingly thorough, especially for races in Hong Kong. Despite the wealth of information available, there did not appear to be an equally large amount of analysis using that data before us, giving us free reign to work on whatever facets of the data intrigued us without being overly concerned about simply retreading someone else’s footsteps.

With the basic concept of our project nailed down and an open track before us, we went about looking for what questions in particular could provide us with an intriguing insight into how the sport of horse racing functions.

## **Research Questions**

Research Question 1; *“Does being a Frontrunner or Slipstreamer result in a better finishing position?”*

Research Question 2; *“Do jockeys affect the performance of horses they ride?”*

Research Question 3; *“Does starting gate offer any advantage to horses in racing?”*

### **R**esearch Question 1: *“Does being a Frontrunner or Slipstreamer result in a better finishing position?”*

In racing, jockeys can often be seen using a variety of different strategies. Holding back and maintaining pace within the pack in an attempt to save energy for the final push or declaring dominance early and maintaining the forward position in the hopes that they will come out victorious. The first research question explores the race strategies involved in the race itself and attempts to find patterns using race data across 1500+ races from Hong Kong. By using data to quantify and potentially show patterns that develop within the race itself, the aim of this question is to see if there is an optimal position to maintain throughout the race that would result in the best possible finishing position overall.

### **R**esearch Question 2: *“Do Jockeys affect the performance of horses they ride?”*

Naturally, a lot of attention is paid to the performance of a horse in professional horse races. This leaves the question however, what impact do the jockeys have? Should we pay attention to who rides what horse, or are the jockeys simply there to successfully steer the horses around the track, having little impact of their own? During the second research question dives into racing data in Hong Kong managed by the South China Morning Post (SCMP) to quantify the performance of individual jockeys compared to individual horses. With this we will be able to determine if jockeys are able to significantly change the finishing position of the horses they ride compared to average, or if those averages remain unmoved.

### **R**esearch Question 3: *“Does starting gate offer any advantage to horses in racing?”*

This question explores the possibility that starting on gate 1 offers an advantage to horses as it is the inside lane. In horse racing starting positions aren’t staggered like in athletics and it would stand to reason to believe that it would offer some form of advantage. By using the same data from research question 1 we can see if there is a stand out advantage and if there is we begin to quantify how much of an advantage exists.

# **Data Wrangling**

## **Data Acquisition**

While looking for data on horse racing we would be able to use, we made the discovery that data from races that take place in Hong Kong were particularly well maintained, even including timings for horses during the course of the race. With this in mind, we decided to constrain the data we would use to Hong Kong horse races, allowing us to ensure our data would be as full as possible and easy to compare. The down side to this strategy however is the fact that we will not be able to compensate for anything that might affect the data in a way unique to Hong Kong compared to the rest of the world, or perhaps even find what affects it in such a way.

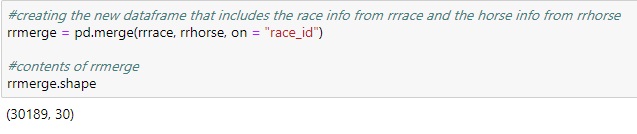
For research questions 1 and 3 data was sourced from Kaggle; A data science website whose goal is to offer content such as this and make it available to the public. The data contains race information of 1,561 local races based in Hong Kong from the 2014 - ‘17 seasons and comprised of two tables which we joined together through the race\_id column. It contains information on race times, finishing positions, sectional positions and horse ids to name a few columns.

The data used for research question two was scraped from HK Racing, a section of the South China Morning Post website that maintains detailed racing data on 1493 horses who have raced in Hong Kong. The data contains race times, which track was used, even the sire and dame of each horse. While all of this was still recorded, research question two focuses on the final race results, the horse performance, and the jockey who was riding them for that race.

## **Data Cleaning & Preparation**

* + 1. **Cleaning Kaggle Data**

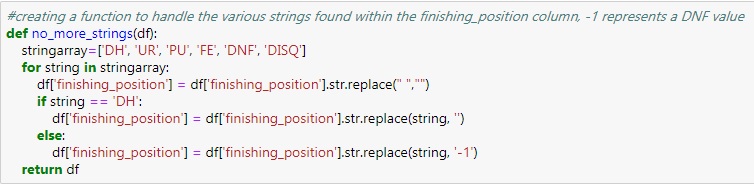
The Kaggle dataset came relatively well cleaned with only a few changes needing to be made for the research questions. There are unique challenges which will be detailed below due to the type of positional data that the dataset offers us. Since we had two tables Race Result Horse (rrhorse) and Race Result Race (rrrace) the first thing that had to be done was join the two tables from the dataset together as certain useful information such as race length and race sectional positions were available on one of the datasets but not on the other. A sample of the code used to merge the two tables together by race\_id is shown below.



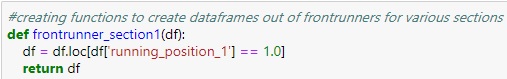
The next challenge we faced when cleaning the Kaggle set was NaN values. Generally NaN values are not a big issue however when dealing with race and sectional data over multiple race lengths, certain races had entire columns of NaN values, an example would be a 1000m race has three sections and the dataset has six columns for each section which will be used for the longer races. This means that 1000m has three columns of usable data and then another three of just NaN values.

Before we progressed we decided that creating individual data frames for each race length would be the most ideal thing to do and given the nature of the research question it would be in the best interest of the research as we can look at patterns specific to race length. After creating the dataframes for each race length we were then able to check and see if any NaN values were still present and if they were we knew that this was a DNF (Did Not Finish) value.

We decided the right way to handle DNFs was to swap their value to -1 rather then just remove them. This will allow us to plot them should we find the need to in the future. The next step was to remove strings that were mixing with the integer values in sectional position data. This was important as any attempt to plot averages into visualisations would cause an error by the presence of strings. Below is an example of the function used to remove strings from the sectional position column.



At this point we considered the dataframe to be clean and ready to answer research questions 1 and 3. However a final challenge with the type of data we were using was that it required a lot of functions to be created for each individual race length due to the different number of columns in each data frame but this was easily handled with said functions. (see example below for function used to handle the creation of frontrunners in section 1 of a race, other functions were created to handle different sections.)



* + 1. **Cleaning SCMP Data**

Despite being manually scraped, the data retrieved from the SCMP website is quite well formed, with many of the column types being properly recognized by Pandas and no obvious breaks caused by NaN values. Naturally however, nothing is perfect. Our first step in cleaning the SCMP data was to check which columns had NaN values in them and display any rows that had NaN values. The resulting data frame contained less than ten rows that luckily only contained NaN values in all of the win\_time and last\_qtr columns.

A brief examination of these rows show that everyone has a ‘DNF’ value in the ‘rank’ column. On a whim we brought up all races that ended in a ‘DNF’ for that horse to find every NaN value as well as a few additional races which apparently did not finish despite having a ‘win\_time’ value. Knowing that all our NaN values are connected to valid DNF values, we will split the data into finished and unfinished races according to the values in ‘rank’ after some more cleaning is complete.

First however, we have to organize the original scrape of each horse’s name. On the SCMP website, the ID, English name, and Chinese name of every horse is contained in the same string so we have to split the recorded string into its appropriate parts while cleaning. We split the original string along the back slash the SCMP uses to separate them and give each part of the horse name its own column. While the Chinese name is immediately dropped, the code still generates it for anyone who may want to use it by simply skipping the cell that drops it for some reason.

We also need to check the data types of all of our data in order to ensure that we’ll be able to properly compare rows to each other. Notably the ‘rank’, ‘horse\_wt’, ‘rt’, ‘odds\_on’, and ‘odds\_last’ were not the integer values they need to be in order to be able to compare different values with methods like greater/less than. After looking at unique values of each column, it can be seen that for all but the ‘rank’ column, Pandas was prevented by auto assigning integer values to these columns due to two blank space characters the SCMP website uses: “\xa0” and “--”. Replacing SCMP blank characters with a uniform blank space of our own allows us to convert the columns to the appropriate numeric types.

Finally, we come back to the ‘rank’ column. The unique values of ‘rank’ range from 1-14, and include string values such as ‘DNF’ and ‘DISQ’. While these string values are not invalid inputs, research question two will only be analyzing data from completed races in order to constrain our analysis to races that were actually finished. Therefore, we’ll remove all rows that do not have a completed rank (remembering this also removes all NaN values), and save them as a separate csv file in the event we want to perform any analysis on incomplete races. This allows us to proceed with our analysis with a now clean dataset, while still preserving everything that was scraped from HKRacing.

# **Data Analysis & Results**

## **Research Question 1:** *“Does being a Frontrunner or Slipstreamer result in a better finishing position?”*

### **Datasets**

The data relevant to this research question consists of the sectional position columns, finishing position columns and race length column. How these columns and the data contained will be used will be discussed in the following section. Finally, the jupyter notebooks associated with this research question is frontrunner\_slipstreamer.ipynb and measuring\_slipstreaming.ipynb.

### **Approach**

Note about definitions: Issues arose when coming up with a way to visualise whether Frontrunners or Slipstreamers were more effective; Our initial definition of Frontrunner was any horse holding a 1st, 2nd or 3rd position in the first two sections of a three section 1000 metre race, however we came to realise that this was assuming too much and not taking into account instances where horses were not in winning positions for the first section but did gain ground for the 2nd.

This issue was addressed by changing our definition of Frontrunner to any horse that is in 1st place only and we would begin to try to visualise this section by section in the races. Our plan was to make a dataframe out of horses in the frontrunner position for section 1, likewise for section 2 the whole way up to whatever amount of sections each race has; This will allow us to get some well rounded data by seeing how those horses finished in their races.

We are defining Frontrunner as any horse holding 1st position at any given point in the race, however defining Slipstreamers was considerably more difficult to nail down; Scientifically slipstreaming is the action of reducing aerodynamic drag by positioning your horse behind or beside another horse. The effect isn't insignificant either, studies show given the right angle it could reduce drag by upwards and above 50%; Here is a link to a web page discussing the topic about wind tunnel simulations done at the Royal Melbourne Institute of Technology;

<https://www.horsejournals.com/slipstreaming-reduces-drag-horse-racing-66-percent>.

We have decided to split slipstreamer into two segments; Slipstreamer and Tailgater. We are defining "Slipstreamer" as any horse that is not in the first 3 positions of a given section to be safe and is holding a position no higher than 8 (position >=4 & <=8), and “Tailgater” as any horse holding a position of > 9. The case could be argued that 3rd place is also slipstreaming, however considering the many races where the first three horses are three abreast we wanted to be safe and am unfortunately limited to accounting any other situations as outliers as the data doesn't have positional data relative to horses and the positions of other horses and jockeys near them.

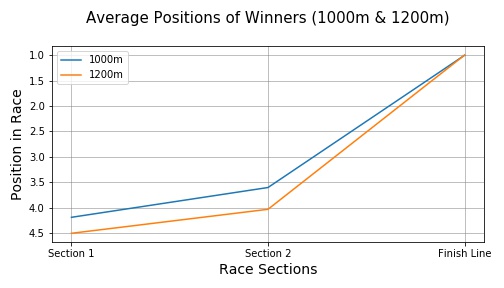
We have decided that the slipstreamer definition will not exceed 8th place in an attempt to prevent dragging the average down needlessly. This allows us to show graphs depicting the effect slipstreaming has at different sections of the pack and not needlessly drag the average down. So, to summarize the definition of Slipstreamer will be any horse in a position of >=4 & <=8 and Tailgaters will be defined as any horse holding a >=9 position.

With these definitions we were able to use the data to plot a variety of visualizations showing the average positions horses would be in section by section and where they would finish. We will be doing this for each race length to see if the length of race also affects the strategies as frontrunner might be a more viable strategy in shorter races given the lesser need to save energy. This will allow us to spot any patterns that might emerge and gain insight into what the best strategy might be and ultimately answer the research question with some well rounded results.

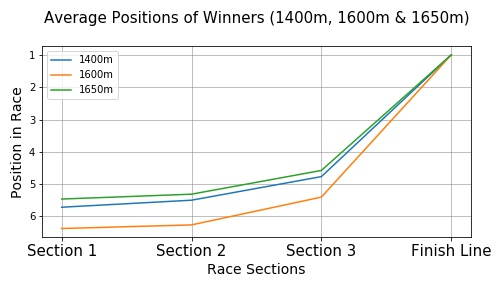
### **Results**

Key findings for this research question were very interesting. We were able to find average positions for every section of each race length and show frontrunner, slipstreamer, tailgater and winner positions throughout each race. Below will begin by taking a look at the winners.

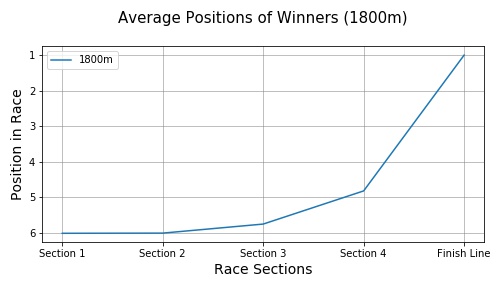
As we can see in figures 1.1, 1.2 and 1.3 below, on average winners of the race are holding a top 6 position at the end of the first section in each race, improving upon that somewhat in the middle sections and then taking the victory in the final section. This suggests that there could be some presence of slipstreaming which we will explore more later. The findings also suggest that on average to win the race horses must be holding a somewhat competitive position coming into the final section for sprint to the finish line. Another interesting trend is the average position of winners at the end of the 1st section of the race seems to fall back the longer the race is.

**

*Figure 1.1 A graph showing average positions of Winners for the given race lengths*

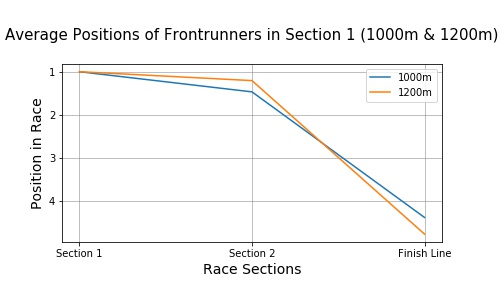
**

*Figure 1.2 A graph showing average positions of Winners for the given race lengths*

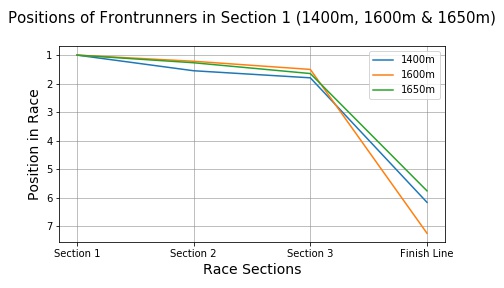
**

*Figure 1.3 A graph showing average positions of Winners for the given race length*

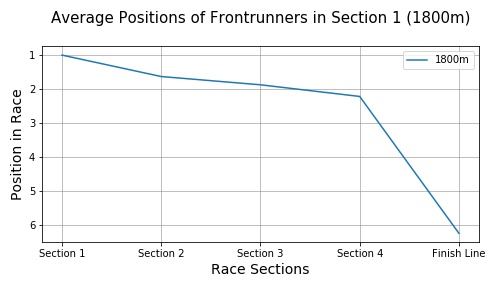
Below, we will take a look at Frontrunners from the races in all lengths and visualise positions throughout each section. My assumption coming into this was that in the shorter length races slipstreaming will have less of an effect throughout and frontrunner would be the preferred race tactic. As we can see in fig 2.1 the average finishing position of horses in 1000m and 1200m is 4th or 5th, when we compare this to the longer races such as figures 2.2 and 2.3 we can see that on average there is a drop from 4th-5th in fig 2.1 to 6th, 7th or even 8th in fig 2.2 and 2.3.



*Figure 2.1 A graph showing average positions of Frontrunners in 1000 & 1200m races*

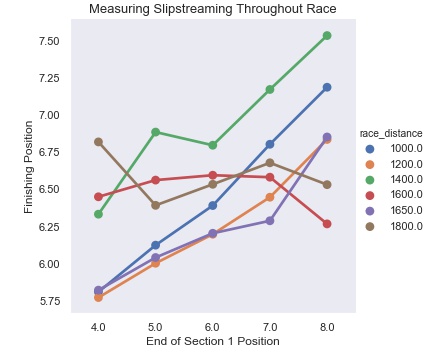
**

*Figure 2.2 A graph showing average positions of Frontrunners in 1400, 1600 & 1650m races*

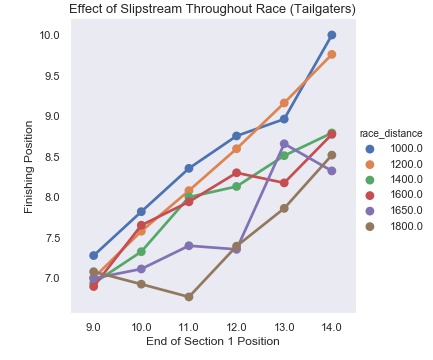
**

*Figure 2.3 A graph showing average positions of Frontrunners in 1800m races*

Next we will take a look at the potential effect slipstreaming has throughout the race for the slipstreamers and the tailgaters. Fig 3.1 and 3.2 below are checking to see how slipstreamers and tailgaters improved their position from section 1 (x-axis) to the end of the race finishing position (y-axis). As we can see in fig 3.1 & 3.2 there is a relatively small if any presence of slipstreaming throughout at any length of race.

**

*Figure 3.1 A graph showing the potential effect of slipstreaming (Slipstreamers)*

**

*Figure 3.2 A graph showing the potential effect of slipstreaming (Tailgaters)*

On average it looks like slipstreaming doesn't have a massive effect across the races we see here with most horses in a top 5 position actually falling back and the further back in the pack we go we see an increase in their position which we feel is due to the fact that the further back in the pack a horse is the more chances there exist to improve.

### **Discussion**

In conclusion, rather than provide a definitive answer as to what race tactic is best between frontrunners and slipstreamers or going to provide the best finishing position, the data shows a few insights that might help in a race. Firstly that frontrunning is more effective on average in shorter races rather than in longer ones. Secondly that in order to win, on average horses need to be somewhere in the top 6 positions coming into the final section in order to have a good shot at victory.

This question has been a rather large example of regression to the mean and we believe it's worthwhile mentioning a few limitations that we have learnt and felt since starting this research question. It's worth noting that using positional data, in theory sounds like a good way to see if slipstreaming had an effect in a race, ultimately with data such as this we find that everything regresses towards the mean.

With what we know now we would attempt to look at the slipstreaming effect from data collected with sensors attached to the horse that can actually measure the aerodynamic drag throughout the race and potentially have gps data that shows their position in relation to other horses in the race. We think it's also worth mentioning that while we feel there's regression to the mean here, the reason why in this context is the simple number of chances provided in the sense that out of the top 4 people, there's only one chance for first place to maintain that first place, and three chances for them to fall behind.

Finally, my answer to the question of *“Does being a frontrunner or slipstreamer provide a better finishing position?”* is rather inconclusive with the data we have, further research needs to be done in order to have the right data to measure slipstreaming in order to calculate how much slipstreaming was done in a race to make a good observation.

With that in mind, we would say with this particular study being a frontrunner is more beneficial as we found that frontrunners of a race would on average drop back to 4th, 5th or 6th by race end however with slipstreamers there was no guarantee that they would make up the ground coming into the final section of the race and therefore it is our opinion that in this case Frontrunners provide the highest chance of finishing in a better position as seen in the tailgaters section above most improvements were only to 5th or 6th place.

## Research Question 2*: “Do jockeys affect the performance of horses they ride?”*

### **Datasets**

The data used for this research question will consist of the cleaned data scraped from the SCMP website, with specific emphasis paid to the ID, Jockey, and Rank columns.

### **Approach**

The first thing we did with the cleaned SCMP data was view the number of races per rank that jockeys or horses have achieved in order to create a baseline for future analysis. While not strictly needed to answer the stated research question, it can be helpful to have a glance at the overall data you are working with in order to mentally orient yourself and help recognize any results later in the analysis that may appear incorrect. (*Figure 4.1*)

Next we drop all but the ID, Jockey, and Rank columns before pivoting the data frame, providing us with a data frame that shows the mean rank jockey ‘x’ achieved while racing horse ‘y’. This pivoted data frame will be the basic structure we use for the large majority of our analysis for research question two. This format can be well visualized with a heat map, (*Figure 4.1.1*) however once you try to show more than a handful of jockeys at a time it quickly becomes unlegible due to the number of variables you are attempting to graph. (*Figure 4.1.2*)

Before beginning the proper analysis of our second research question, we split our data according to ‘pro’ and ‘new’ jockeys in order to ensure that we could show all available data without having possible outliers affect more reliable data. For our results, we decided a ‘new’ jockey was any jockey who has only one to nine recorded races, while ‘pro’ jockeys consist of any jockey that has ten or more recorded races.

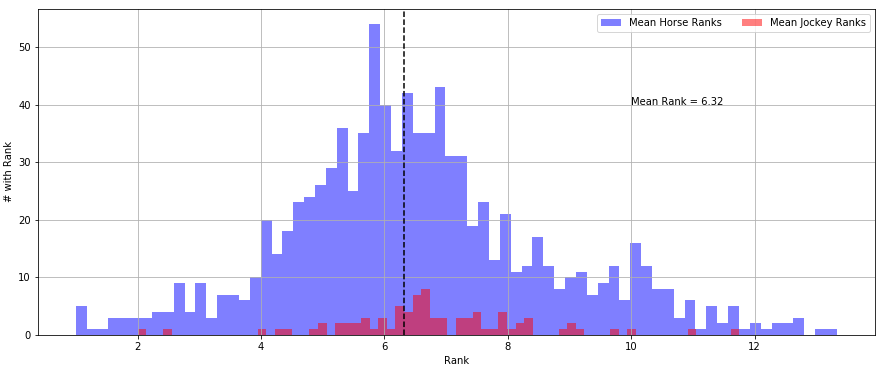
With the data split, we normalize the data, giving every jockey a z-score when compared to the average and standard deviation of all jockeys recorded. Graphing this comparison as a histogram and displaying the mean of each split allows us to quantify the difference in quality between experienced and fresh jockeys. (*Figure 4.2*) The results we recorded show a difference between ‘pro’ and ‘new’ ranks of almost a full standard deviation, implying that jockeys we have more data recorded for perform better than ones who have raced less. While not exactly a surprising result, it is good to confirm any assumptions one might make about our data before proceeding, and gives us good reason to focus only on ‘pro’ jockeys to allow as much of a chance for a jockey to have an impact as possible.

With our split data normalized now, we calculate means and z-scores for every horse in our data as well in order to be able to compare the impact a jockey has on that horse’s performance. The z-score of each horse is given a graph as well to allow the reader to get an idea of what the base performances of horse’s in this data set are. (*Figure 4.2.1*)

Finally, we are able to get to the crux of research question two. In a new data frame, we give every jockey a z-score relative to the horses they race. This is done by calculating a z-score for every time a jockey rides a horse by using that horse’s mean and standard deviation to show which jockeys perform above the average and by how much. Once this is done for the entire pivot structure we created, we average the z-score each jockey got with all the horses they raced and graph them using the same histograms we have been working with for the majority of this research question. (*Figure 4.3*) The result showed, much to our surprise, that while jockeys are capable of making a significant impact on the horses they race, the large majority of jockeys have little to no impact whatsoever on the performance of the horses they race.

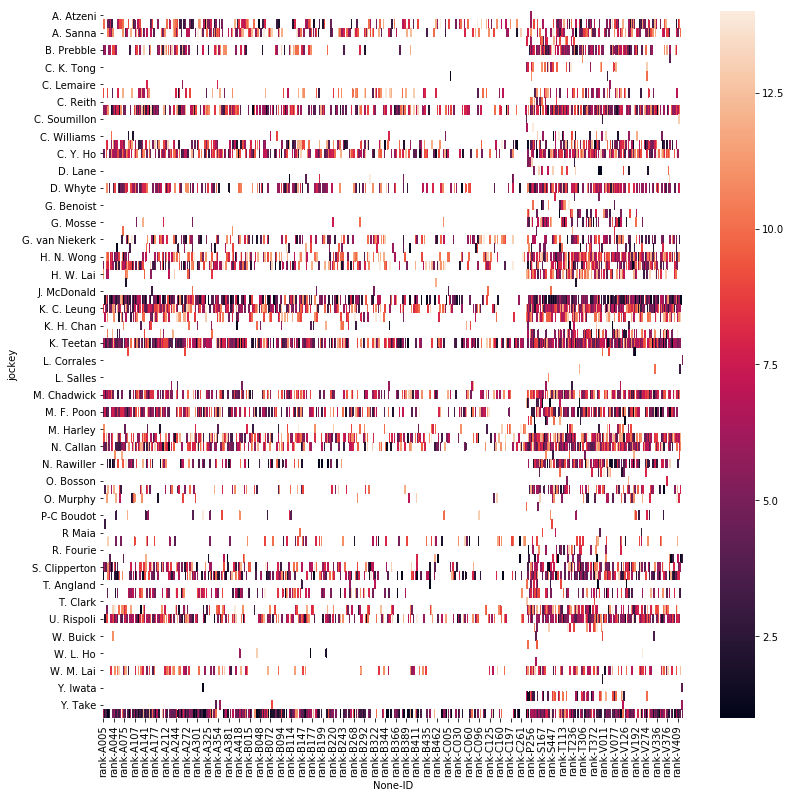
In order to ensure this isn’t a fluke result, we graphed jockeys who have only one hundred or more races recorded. (*Figure 4.4*) Even when restricted to some of the most experienced jockeys on record, the mean impact was almost zero. As one final check to give jockeys as good of a chance as physically possible to make a large impact on average, we created data frames using only ‘pro’ jockeys. From there, we restrict the results to only when ‘good’ jockeys (judged as a z-score of above zero) race ‘bad’ horses (judged as a z-score below zero) versus when they race ‘good’ horses (judged as z-score of greater than or equal to zero). (*Figure 4.4.1*) Even in this optimum position, where jockeys have the most room to make an improvement or where they have the most to work with respectively, the mean impact the jockeys were able to make is still almost exactly zero.

### **Results**

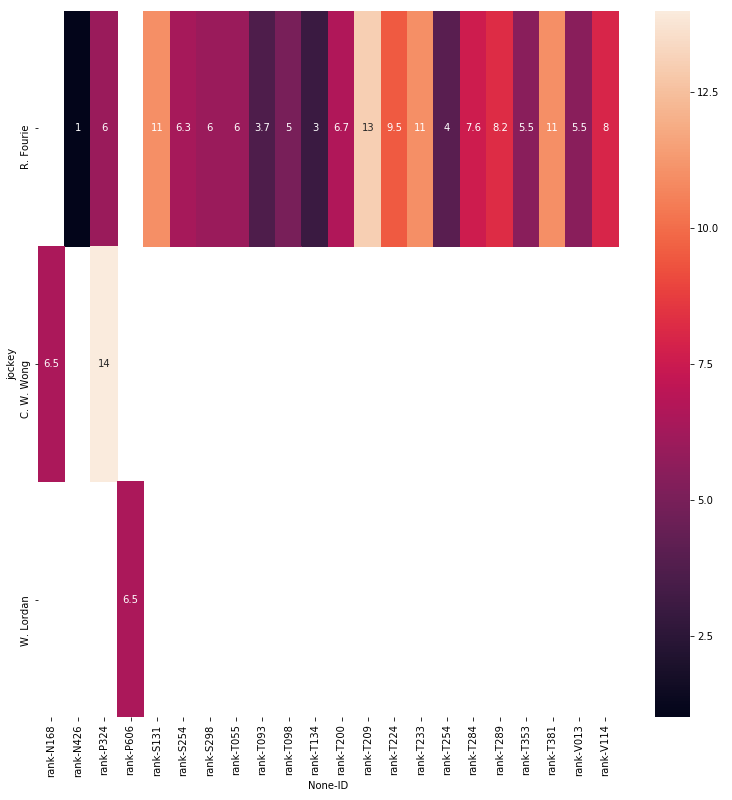


*Figure 4.1 This graph shows the number of all ranks achieved in Research Question 2 and displays a normal distribution, which is to be expected for a random population.*

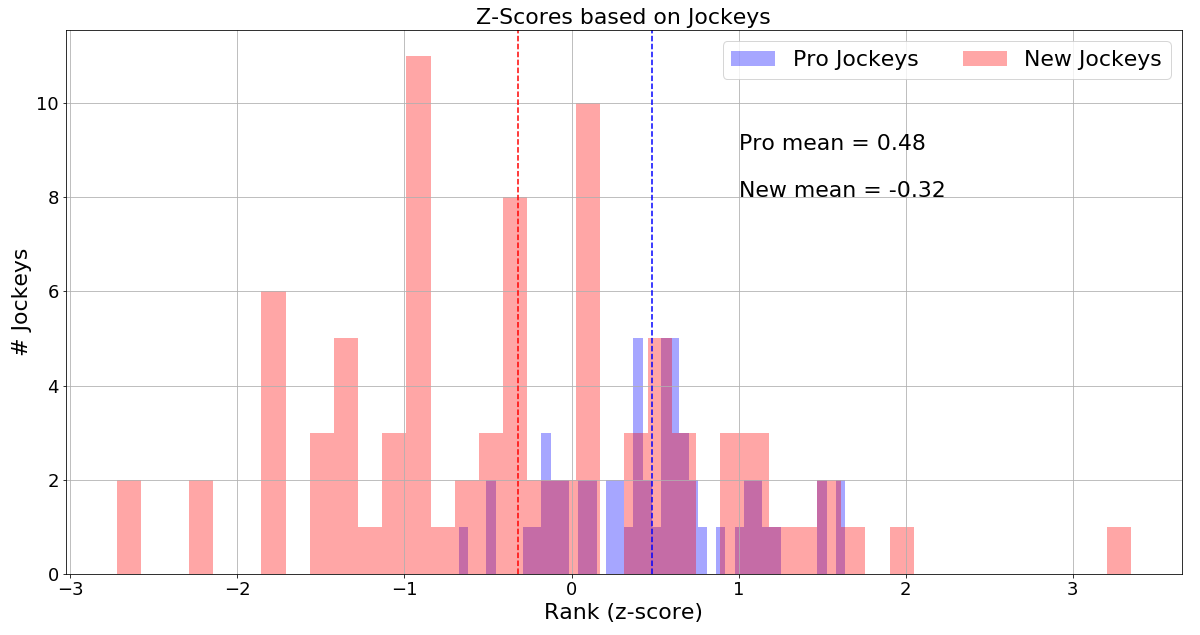
asdf



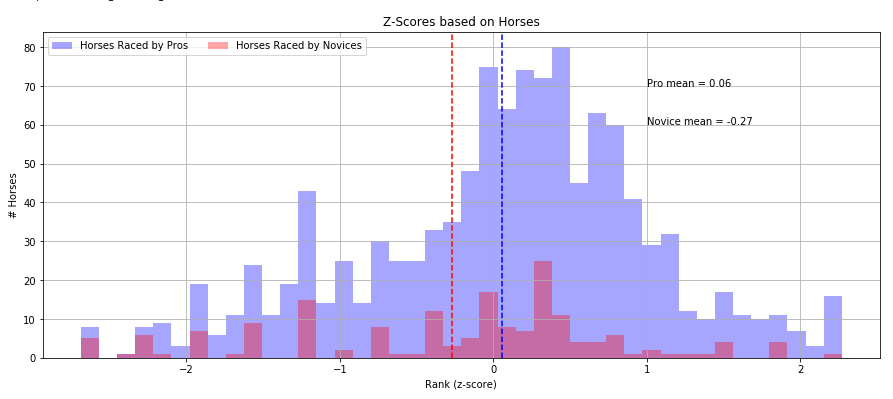
*Figure 4.1.1 A heatmap attempting to display all rank data from the SCMP, however the size makes it difficult to read.*



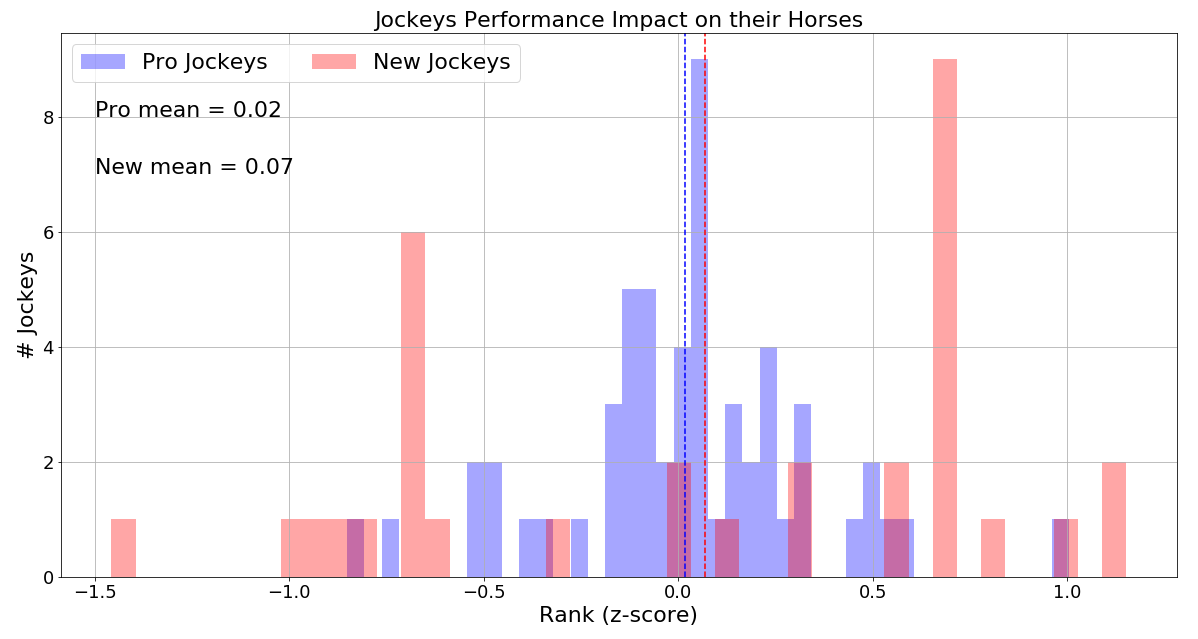
*Figure 4.1.2 A smaller heatmap displaying a sample of the data shown above that is legible.*



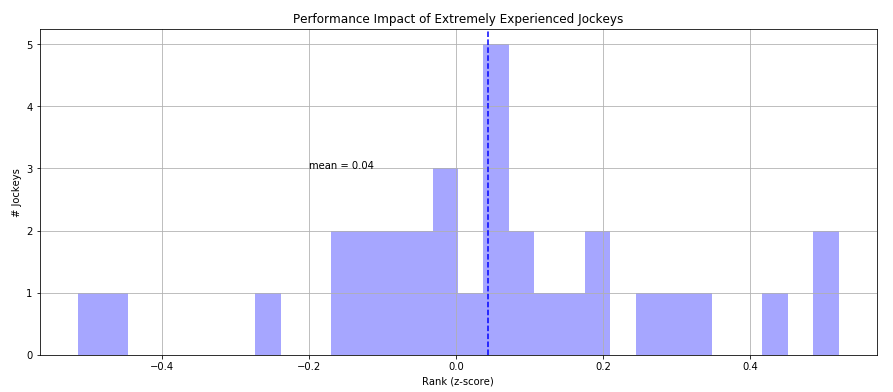
*Figure 4.2 A histogram comparing the z-scores of ‘pro’ jockeys and ‘new’ jockeys.*



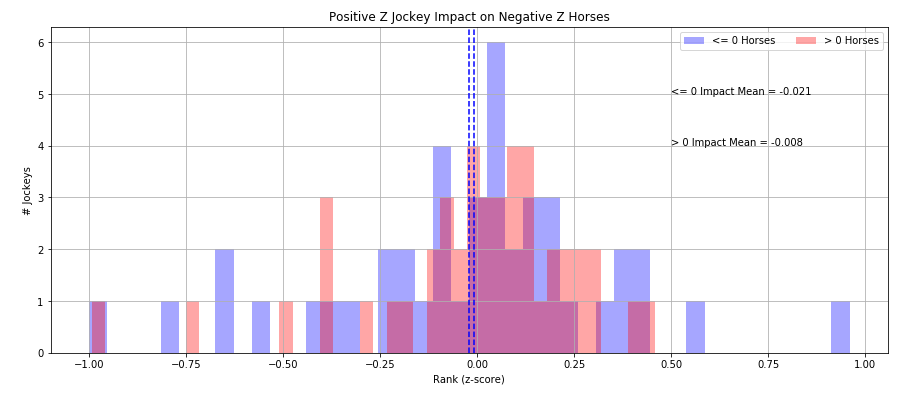
*Figure 4.2.1 A histogram comparing the z-scores of horses raced by ‘pro’ jockeys and horses raced by ‘new’ jockeys.*



*Figure 4.3 The quintessential visualization of research question two showing the impact ‘pro’ and ‘new’ jockeys can have on the performance of the horses they ride.*



*Figure 4.4 A restriction of Figure 4.3, showing only jockeys who have at minimum one hundred races on record.*



*Figure 4.4.1 A histogram showing the impact ‘pro’ jockeys have on ‘bad’ horses compared to the impact they have on ‘good’ horses.*

### **Discussion**

In conclusion, we were able to prove that jockeys are capable of improving the performance of the horses. There are examples of jockeys in our data who on average get at least one rank above their horse’s mean performance. That being said the large majority of jockeys are unable to make *any* notable impact, positive or negative, to the average rank the horses they race are able to achieve. Ultimately this means that while the jockey racing a horse is still a factor that should absolutely be taken into account, it *usually* does not have an impact.

## **Research Question 3:** *“Does starting gate offer any advantage to horses in racing?”*

### **Datasets**

The data relevant to this research question consists of the draw columns and finishing position columns. The Draw column contains data on what gate the horses started in for each race, it’s called “draw” due to the random number the horse drew a few days before the race. How these columns and the data contained will be used will be discussed in the following section. Finally, the jupyter notebook associated with this research question is draw\_testing.ipynb

### **Approach**

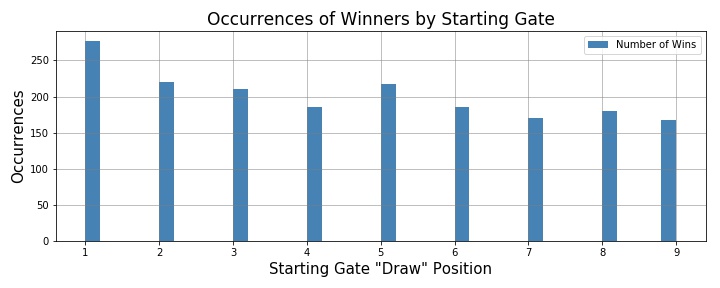
For this question, my approach was to create a dataframe that had the draw column and all the 1st place finish positions horses. This allows us to plot a visualisation counting up the number of occurrences of 1st place wins and the starting gate they are associated with. My assumption coming into this question would be that in some way the inside lane (gate 1) would provide a small advantage of some kind. There were no definitions needed for this question.

### **Results**

Below are two visualisations that show the number of occurrences of first place wins relative to starting gate of the horse. Fig 4.1 shows all possible gates and the implot line (red line above) shows the how many races contained that number of horses. Put simply, this implot line tells us is that there are considerably less races with fourteen horses present in them then say, five horses. However the implot line also tells us that there are a relatively equal number of horses present in races up to nine horses. Fig 4.2 is be the adjusted visualisation that takes into account this potentially misleading parameter and only shows the balanced entries which give us a much better idea of the disparity between the start gates.

### 

*Figure 5.1 A graph showing number of 1st place wins by starting gate*

**

*Figure 5.2 The adjusted graph showing number of 1st place wins by starting gate*

### **Discussion**

In conclusion, We believe the findings of this question are quite interesting. After adjusting the data it's quite telling how much of a tangible advantage there exists in just getting an inside lane start gate. Clearly this is not so much of an advantage that being assigned another gate is a roadblock to finishing the race in first place however there is enough of a difference that exists in this dataset to show a distinct advantage to having an inside lane. It puts into question why horse racing doesn’t stagger the start gates to cater for this almost obvious advantage from the inside lane.

# **Discussion**

## **Ethical Considerations**

Thankfully we couldn’t think of a way that our findings could be misused or exploited and that is one of the advantages to doing a project with this subject matter. With that in mind, while scraping data from HK Racing, we made sure to read the terms and conditions to see if there were any issues in regards to taking the data. HK Racing however did not have any issues with scraping so long as we didn’t take their code.

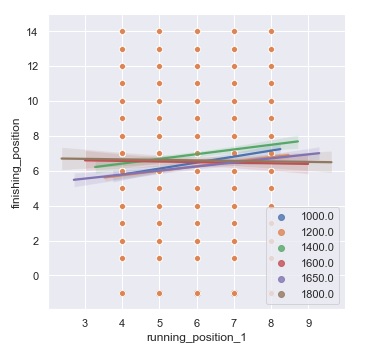
## **Reproducibility**

In an attempt to make sure our results are as easy to reproduce as possible, we made sure to document our code with markdown cells as well as possible to allow anyone else to follow our logic. A specification file is also included with our code to ensure that anyone who would like to run our code for themselves is able to do so in an identical environment. If you have any questions or concerns, please feel free to contact us by emailing: [djowen777@gmail.com](mailto:djowen777@gmail.com) or [andrewbmcgurk@gmail.com](mailto:andrewbmcgurk@gmail.com).

## **Limitations**

Certain limitations existed in our ability to fully measure slipstreaming. As a scientific concept slipstreaming is the act of reducing ones aerodynamic drag by your position relative to another. Coming into this project we feel my work has highlighted the importance of gathering data more suited to measuring the slipstream effect.

Another limitation was the makeup of the data itself. When attempting initially to plot slipstreamers data to a scatter plot to show the potential slipstream effect the plot would show a box full of markers (see below Fig 5.1) Due to the positional data, since there only exists positions 1 - 14 at most, In a sample size of 1000 races every occurrence would be showing on the scatter plot and therefore they were not optimal for the job, we did however use factor plots like the ones you see above.



*Figure 6.1 An example of limited scatter plots showing every occurance*

Our final limitation was simply the source of our data. Because we constrained ourselves to only data recorded in Hong Kong due to ease and completeness, any variables that exist only in horse racing in Hong Kong would affect all of our data and could be very difficult to spot. In fact it is known that Hong Kong racing takes some steps to attempt to make the races as competitive as possible. Despite this

# **Conclusions & Future Work**

In conclusion, as a team we set out to examine certain race tactics and factors that affect racing under a microscope and we feel we achieved that. While some of our findings ended up being considerably difficult to quantify we feel that others provided some very telling results and deepened our knowledge of the topics discussed above.

In regards to the slipstream effect, it would be a great opportunity in future to revisit this question with data collected from sensors placed on the horses themselves. While our findings were difficult to quantify, we feel that they have been beneficial in helping us understand what the slipstream effect actually is and what needs to be done in order to measure it correctly. Also we do not discount the findings from the Frontrunner v Slipstreamer question entirely, interesting trends were shown from the frontrunner perspective aswell as how horses further in the pack generally behave throughout a race. However, in future we feel it would be great to return to unfinished business in this regard.

Furthermore, we believe that the horse racing authority needs to take the advantage of the inner lane seriously within their sport and make more of an effort to balance the races with some kind of staggered start line.

It would also be interesting to work with some of the other pieces of data we recorded given enough time, doing analysis on things such as the effect weight has on a horse’s performance or certain blood lines of horses that are particularly good on the racetrack. Given time, there is still a lot of work that could be done with the data we gathered during the course of this project.

# **Bibliography**

Hong Kong Horse Racing Dataset. Retrieved March ‘19 from: <https://www.kaggle.com/lantanacamara/hong-kong-horse-racing>

HK Racing Dataset Scraping. Retrieved March ‘19 from: <https://www.scmp.com/sport/racing>

Slipstream Effect Study, Carried out by Royal Melbourne Institute of Technology, Retrieved April ‘19 from: <https://www.horsejournals.com/slipstreaming-reduces-drag-horse-racing-66-percent> & <https://phys.org/news/2014-11-horse-position-cent.html>

1. See https://www.cs.ucd.ie/sites/default/files/cs-plagiarism-policy\_august2017.pdf [↑](#footnote-ref-0)