Data_Wrangling

July 31, 2018

1 Data Cleaning and Processing

In this notebook, we will extract the useful variables from the data for building the prediction model in next notebook.

First, we import the dataset that provided by Kaggle Dataset-Can You Predict The Result? . The dataset contains the race result of 1561 local races throughout Hong Kong racing seasons 2014-16 and more information will be added into the dataset. Also, we need to download the race result csv file since that csv file contains date informations of all races and we need to use it as one of our variables.

1.0.1 Import both csv files and merge them by race_id columns

Out[4]:

	src	race_date	race_course	race_number	race_id	race_class	\
0	20140914-1.html	2014-09-14	Sha Tin	1	2014-001	Class 5	
1	20140914-10.html	2014-09-14	Sha Tin	10	2014-010	Class 2	
2	20140914-2.html	2014-09-14	Sha Tin	2	2014-002	Class 5	
3	20140914-3.html	2014-09-14	Sha Tin	3	2014-003	Class 1	
4	20140914-4.html	2014-09-14	Sha Tin	4	2014-004	Class 4	

	race_distance	track_condition	race_name	\
0	1400	GOOD TO FIRM	TIM WA HANDICAP	
1	1400	GOOD TO FIRM	COTTON TREE HANDICAP	
2	1200	GOOD TO FIRM	TIM MEI HANDICAP	
3	1200	GOOD TO FIRM	THE HKSAR CHIEF EXECUTIVE'S CUP (HANDICAP)	
4	1200	GOOD TO FIRM	LUNG WUI HANDICAP	

	track	sectional_time	\
0	TURF - "A" COURSE	13.59 22.08 23.11 23.55	
1	TURF - "A" COURSE	13.55 22.25 22.89 22.85	
2	TURF - "A" COURSE	24.06 22.25 23.66	
3	TURF - "A" COURSE	23.42 22.48 22.47	
4	TURF - "A" COURSE	24.00 22.62 22.64	

		incident_report
0	\n	When about to enter the trac...
1	\n	SMART MAN was slow to begin...
2	\n	ALLEY-OOP and FLYING KEEPER ...

```
3  \n                On arrival at the Start, it ...
4  \n                Just prior to the start bein...
```

Out[5]:

	finishing_position	horse_number	horse_name	horse_id	jockey	trainer	actual_weight	declared_horse_weight	draw	length_behind_winner	...	running_position_3	running_position_4	finish_time	win_odds	running_position_5	running_position_6	race_id	race_date	race_distance	race_class
0	1	1.0	DOUBLE DRAGON	K019	B Prebble	D Cruz	133	1032	1	-	...	2.0	1.0	1.22.33	3.8	NaN	NaN	2014-001	2014-09-14	1400	Class 5
1	2	2.0	PLAIN BLUE BANNER	S070	D Whyte	D E Ferraris	133	1075	13	2	...	9.0	2.0	1.22.65	8	NaN	NaN	2014-001	2014-09-14	1400	Class 5
2	3	10.0	GOLDWEAVER	P072	Y T Cheng	Y S Tsui	121	1065	3	2	...	1.0	3.0	1.22.66	5.7	NaN	NaN	2014-001	2014-09-14	1400	Class 5
3	4	3.0	SUPREME PROFIT	P230	J Moreira	C S Shum	132	1222	2	2	...	5.0	4.0	1.22.66	6.1	NaN	NaN	2014-001	2014-09-14	1400	Class 5
4	5	7.0	THE ONLY KID	H173	Z Purton	K W Lui	125	1136	9	4-1/4	...	10.0	5.0	1.23.02	6.1	NaN	NaN	2014-001	2014-09-14	1400	Class 5

[5 rows x 22 columns]

1.0.2 Remove the unused columns - running_position , length_behind_winner

Running positions are indicators of running style of horses, but there are so many missing values of the runnning positions in the dataset. Also, there are so many missing values of the length behind winner column. Therefore we will not use them for the prediction and analysis.

Out[7]:

	finishing_position	horse_number	horse_name	horse_id	jockey
0	1	1.0	DOUBLE DRAGON	K019	B Prebble

```
1                  2           2.0  PLAIN BLUE BANNER     S070    D Whyte
2                  3          10.0         GOLDWEAVER     P072  Y T Cheng
3                  4           3.0     SUPREME PROFIT     P230  J Moreira
4                  5           7.0       THE ONLY KID     H173   Z Purton

        trainer actual_weight declared_horse_weight draw finish_time win_odds  \
0        D Cruz           133                  1032    1     1.22.33      3.8
1  D E Ferraris           133                  1075   13     1.22.65        8
2      Y S Tsui           121                  1065    3     1.22.66      5.7
3      C S Shum           132                  1222    2     1.22.66      6.1
4       K W Lui           125                  1136    9     1.23.02      6.1

    race_id   race_date  race_distance race_class
0  2014-001  2014-09-14           1400    Class 5
1  2014-001  2014-09-14           1400    Class 5
2  2014-001  2014-09-14           1400    Class 5
3  2014-001  2014-09-14           1400    Class 5
4  2014-001  2014-09-14           1400    Class 5
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30189 entries, 0 to 30188
Data columns (total 15 columns):
finishing_position       30187 non-null object
horse_number             29851 non-null float64
horse_name               30189 non-null object
horse_id                 30189 non-null object
jockey                   30189 non-null object
trainer                  30189 non-null object
actual_weight            30189 non-null object
declared_horse_weight    30189 non-null object
draw                     30189 non-null object
finish_time              30189 non-null object
win_odds                 30189 non-null object
race_id                  30189 non-null object
race_date                30189 non-null object
race_distance            30189 non-null int64
race_class               30189 non-null object
dtypes: float64(1), int64(1), object(13)
memory usage: 3.7+ MB
```

1.0.3 Remove unused rows and data

Some finishing positions are special incidents, such as,**WV**, **WV-A**, etc. Please refer to this page for the descriptions. Thus, we want to remove the finishing positions which are not numbers.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 29364 entries, 0 to 30186
```

```
Data columns (total 15 columns):
finishing_position       29364 non-null float64
horse_number             29364 non-null float64
horse_name               29364 non-null object
horse_id                 29364 non-null object
jockey                   29364 non-null object
trainer                  29364 non-null object
actual_weight            29364 non-null float64
declared_horse_weight    29364 non-null float64
draw                     29364 non-null float64
finish_time              29364 non-null object
win_odds                 29364 non-null float64
race_id                  29364 non-null object
race_date                29364 non-null object
race_distance            29364 non-null int64
race_class               29364 non-null object
dtypes: float64(6), int64(1), object(8)
memory usage: 3.6+ MB
```

1.0.4 Convert finishing position to 1/0 (Winner = 1)

Since we only want to predict which horse is the winner and find the winning probability, we need to convert finishing position column to 1/0

1.1 Extract information from the dataset and convert to predictive variables

1.1.1 Convert Date from factor to date format

1.1.2 Convert finish time from string to integer (measure in second)

```
Out[14]: 0    82.33
         1    82.65
         2    82.66
         3    82.66
         4    83.02
         Name: finish_time, dtype: float64
```

1.1.3 Find the speed of the horse

since each race has a different distance, it's unfair to determine the speed of a horse by the finish time. We can find the real speed by dividing the finish time by distance.

1.1.4 Jockey Statisitc(the winning percentage of jockey)

Count the number of time of each finishing position for every jockey. And calculate the percentage of each finishing position.

```
Out[17]:    finishing_position  horse_number     horse_name horse_id     jockey  \
         0                 1.0           1.0  DOUBLE DRAGON     K019  B Prebble
```

```
1                11.0           7.0    AUTUMN GOLD     P044  B Prebble
2                13.0           4.0   EXAGGERATION     S226  B Prebble
3                11.0           8.0     BEST TANGO     S121  B Prebble
4                 8.0           7.0  CULTURAL CITY     N263  B Prebble

   trainer  actual_weight  declared_horse_weight  draw  finish_time  win_odds  \
0   D Cruz          133.0                 1032.0   1.0        82.33       3.8
1  S Woods          123.0                 1011.0  14.0        82.34      21.0
2  J Moore          127.0                 1141.0   4.0        57.74      57.0
3   W Y So          123.0                 1089.0   2.0        82.78       8.0
4   W Y So          124.0                 1070.0   9.0        83.64      41.0

    race_id  race_date  race_distance race_class  win      speed  jockey_1st  \
0  2014-001 2014-09-14           1400    Class 5    1  17.004737    9.421365
1  2014-010 2014-09-14           1400    Class 2    0  17.002672    9.421365
2  2014-005 2014-09-14           1000    Class 4    0  17.319016    9.421365
3  2014-006 2014-09-14           1400    Class 3    0  16.912298    9.421365
4  2014-007 2014-09-14           1400    Class 4    0  16.738403    9.421365

   jockey_2nd  jockey_3rd
0   10.905045   10.163205
1   10.905045   10.163205
2   10.905045   10.163205
3   10.905045   10.163205
4   10.905045   10.163205
```

1.1.5 Trainer Statisitc(the winning percentage of Trainer)

Count the number of time of each finishing position for every Trainer. And calculate the percentage of each finishing position.

Out[19]:
```
   finishing_position  horse_number      horse_name horse_id     jockey  \
0                 1.0           1.0   DOUBLE DRAGON     K019  B Prebble
1                 4.0           5.0        SPURS ON     N428  B Prebble
2                 7.0          12.0  HOLLYWOOD KISS     M126  B Prebble
3                 3.0           4.0         MAC ROW     N252  B Prebble
4                 7.0           8.0  HOLLYWOOD KISS     M126  B Prebble

  trainer  actual_weight  declared_horse_weight  draw  finish_time  \
0  D Cruz          133.0                 1032.0   1.0        82.33
1  D Cruz          130.0                 1043.0   1.0        70.36
2  D Cruz          120.0                 1053.0   2.0        83.99
3  D Cruz          131.0                 1000.0   9.0        83.26
4  D Cruz          122.0                 1047.0   4.0       113.08

     ...       race_distance race_class win      speed jockey_1st  \
0    ...                1400    Class 5   1  17.004737   9.421365
1    ...                1200    Class 5   0  17.055145   9.421365
```

```
2     ...               1400    Class 5   0  16.668651   9.421365
3     ...               1400    Class 4   0  16.814797   9.421365
4     ...               1800    Class 5   0  15.917934   9.421365

   jockey_2nd  jockey_3rd  trainer_1st  trainer_2nd  trainer_3rd
0   10.905045   10.163205     4.733132     6.545821     7.653575
1   10.905045   10.163205     4.733132     6.545821     7.653575
2   10.905045   10.163205     4.733132     6.545821     7.653575
3   10.905045   10.163205     4.733132     6.545821     7.653575
4   10.905045   10.163205     4.733132     6.545821     7.653575

[5 rows x 23 columns]
```

1.1.6 DaySince(Number of Days since the last race)

Split the dataset by the horse_id and calculate the number of days since the last race. This variable is an indicator of whether the horse has enough rest. It cannot be directly seen from the data. We can find this variable through some calculation.

Out[21]:

```
   finishing_position  horse_number  horse_name horse_id     jockey  trainer  \
0                11.0           6.0  BURST AWAY     A001    G Mosse  K L Man
1                11.0           6.0  BURST AWAY     A001  M L Yeung  K L Man
2                 6.0           8.0  BURST AWAY     A001    G Mosse  K L Man
3                 2.0           6.0  PRAWN BABA     A002  J Moreira   J Size
4                 2.0           2.0  PRAWN BABA     A002   Z Purton   J Size

   actual_weight  declared_horse_weight  draw  finish_time    ...     \
0          125.0                 1083.0  13.0        70.04    ...
1          124.0                 1073.0   6.0        71.86    ...
2          124.0                 1054.0   1.0        70.25    ...
3          125.0                 1101.0   3.0        95.07    ...
4          130.0                 1096.0   7.0        94.39    ...

   race_class win      speed  jockey_1st jockey_2nd  jockey_3rd  trainer_1st  \
0     Class 3   0  17.133067    8.523592   8.219178    9.589041     6.757783
1     Class 3   0  16.699137    5.263158   5.623648    6.560923     6.757783
2     Class 3   0  17.081851    8.523592   8.219178    9.589041     6.757783
3     Class 3   0  16.829704   24.691992  15.092402   12.782341    15.342466
4     Class 3   0  16.950948   17.442582  14.400993    9.310987    15.342466

   trainer_2nd  trainer_3rd  daysince
0     6.529992     6.757783       0.0
1     6.529992     6.757783      24.0
2     6.529992     6.757783      26.0
3    12.808219     9.863014       0.0
4    12.808219     9.863014      15.0

[5 rows x 24 columns]
```

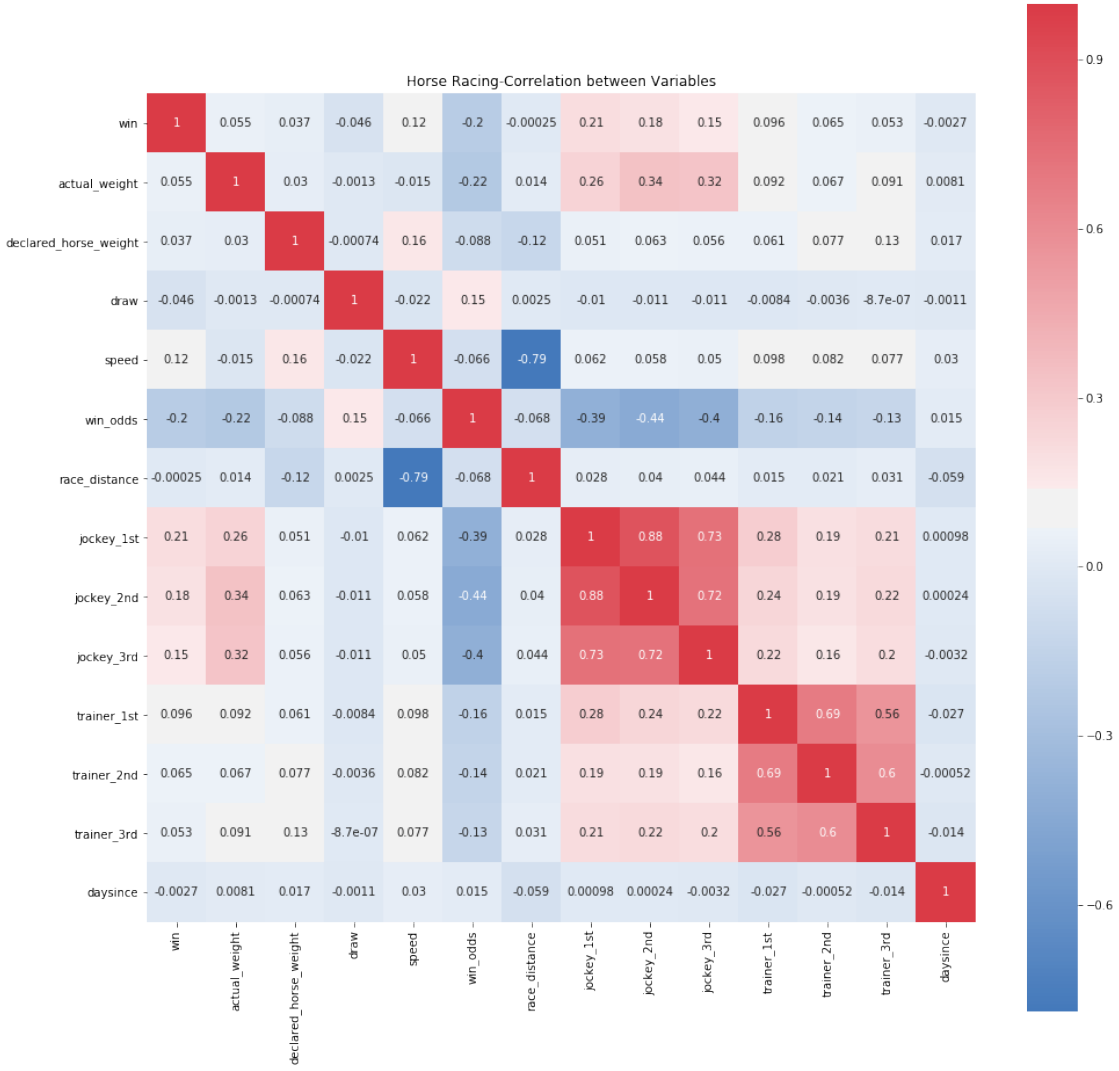
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29364 entries, 0 to 29363
Data columns (total 24 columns):
finishing_position       29364 non-null float64
horse_number             29364 non-null float64
horse_name               29364 non-null object
horse_id                 29364 non-null object
jockey                   29364 non-null object
trainer                  29364 non-null object
actual_weight            29364 non-null float64
declared_horse_weight    29364 non-null float64
draw                     29364 non-null float64
finish_time              29364 non-null float64
win_odds                 29364 non-null float64
race_id                  29364 non-null object
race_date                29364 non-null datetime64[ns]
race_distance            29364 non-null int64
race_class               29364 non-null object
win                      29364 non-null int64
speed                    29364 non-null float64
jockey_1st               29364 non-null float64
jockey_2nd               29364 non-null float64
jockey_3rd               29364 non-null float64
trainer_1st              29364 non-null float64
trainer_2nd              29364 non-null float64
trainer_3rd              29364 non-null float64
daysince                 29364 non-null float64
dtypes: datetime64[ns](1), float64(15), int64(2), object(6)
memory usage: 5.4+ MB
```

1.2 Exploratory data analysis

Finishing_position is the independent variable. **Actual_weight, declared_horse_weight, draw, finish_time, win_odds, race_class, race_distance, jockey_1st, jockey_2nd, jockey_3rd, trainer_1st, trainer_2nd, trainer_3rd, daysince** are the dependent variables for the prediction model.

1.2.1 Correlation Map - correlation between Variables

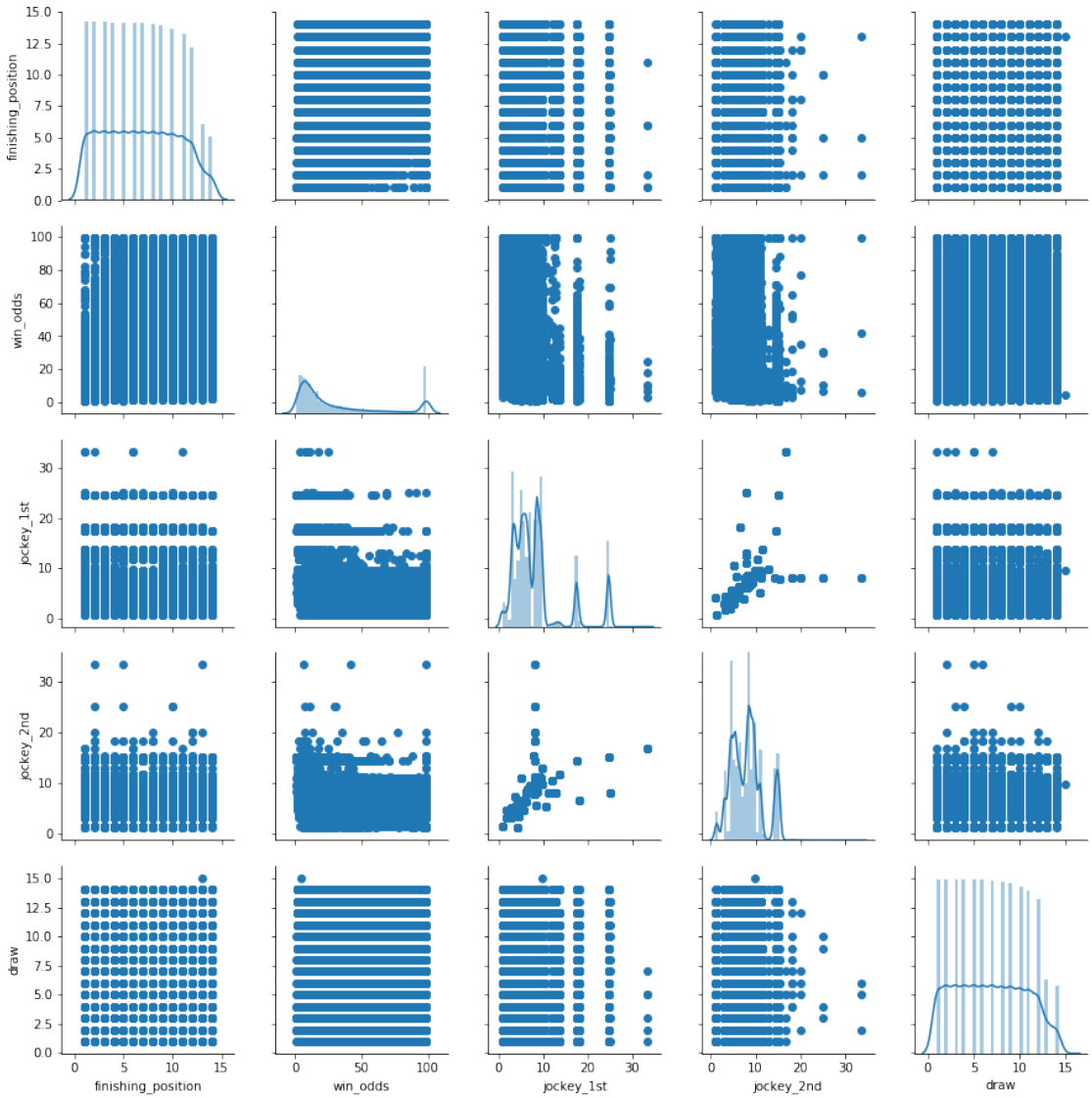
We can find the relationship between variables by plotting the correlation map.



We found that the correlation coefficient between race distance and finish time is 1. When correlation coefficient is 1, it implies that race distance and finish time have a perfect positive relationship. It makes sense because it takes more time for a longer distance, and vice versa.

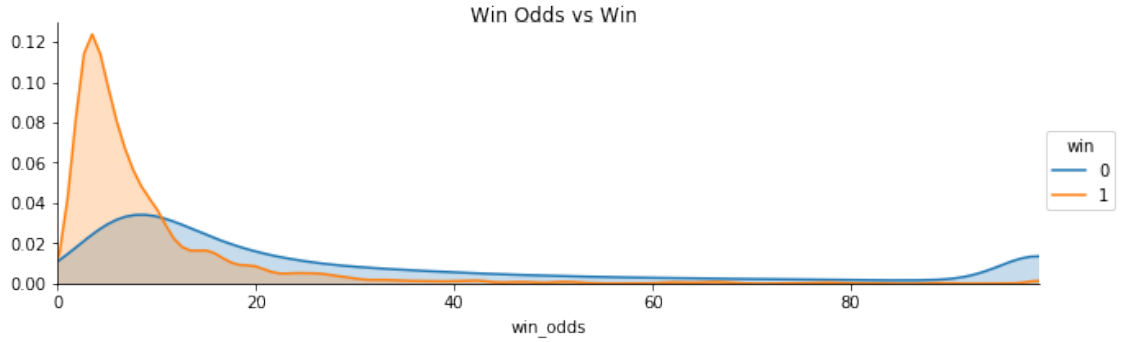
As we can see the win_odds, jockey_1st, speed and draw have higher negative/positive correlations(0.48, -0.28, -0.27, 0.13, 0.096) with the finishing_position. We can plot the plotting pairwise relationships between these variables.

1.2.2 Pairwise Relationships Between Variables



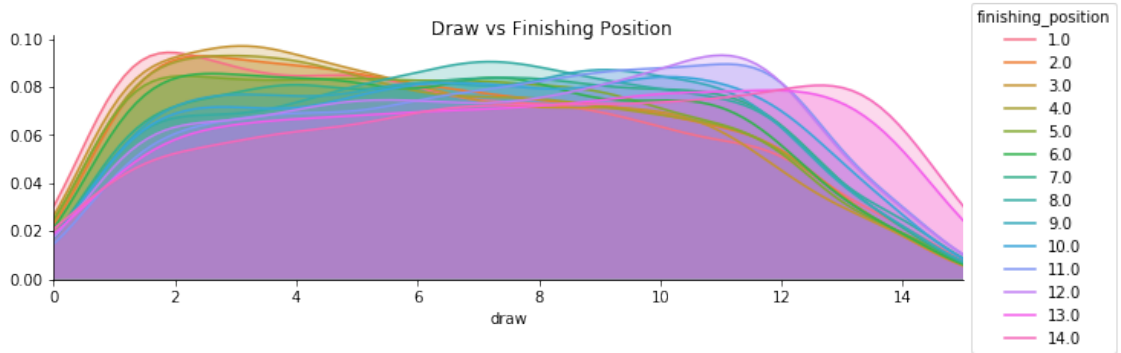
1.2.3 Separate Plots and Analysis of High Correlations Variables vs Finishing Position and Win (Finishing Position =1)

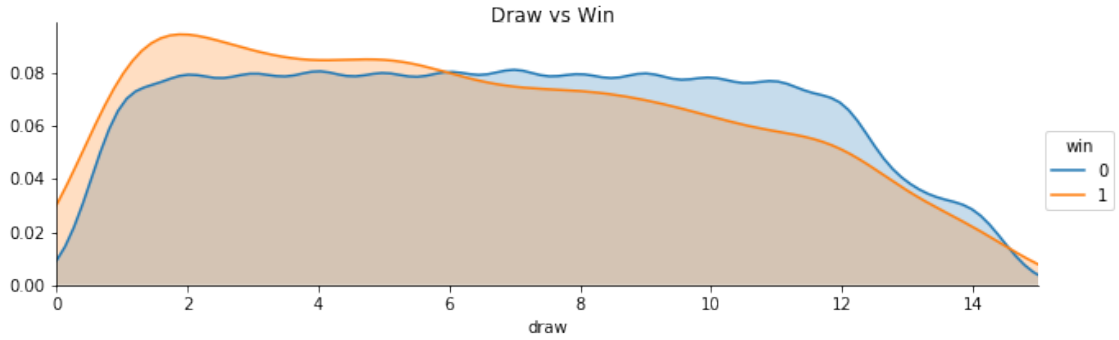
Win Odds vs Finishing Position (Win)



As we can see from the graphs above, it shows that win odds less than 15 have higher rates of a good finishing position, especially when finishing positions 1. It implies the lower the win odds, the higher the chane of winning.

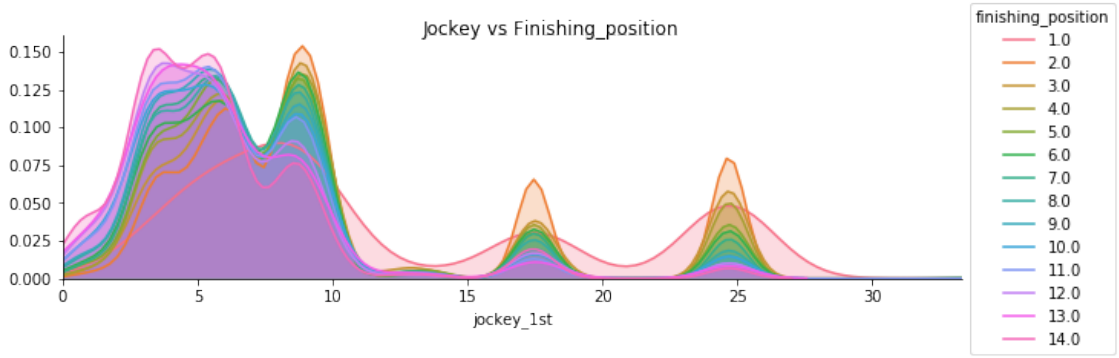
Draw vs Finishing Position (Win) Draw of a horse decides in which individual stall a horse is placed. The smaller the draw number, the closer the horse is to the inside rail, it means smaller draw numbers has an slight advantage over larger draw numbers since a shorter distance to be covered at the turns.

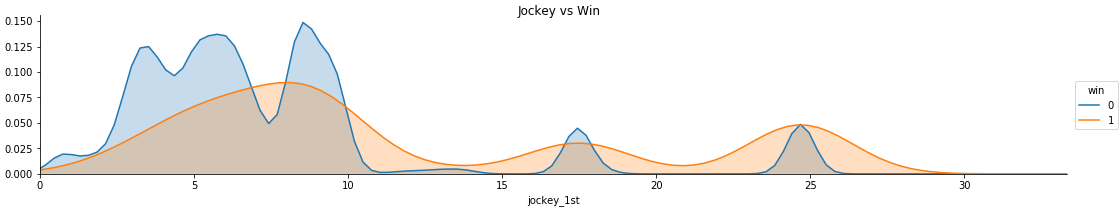




As we can see from the graphs above, it is proved that draw numbers smaller than 6 has an advantage over draw numbers larger than 6. And the smaller the draw number, the larger the advantage.

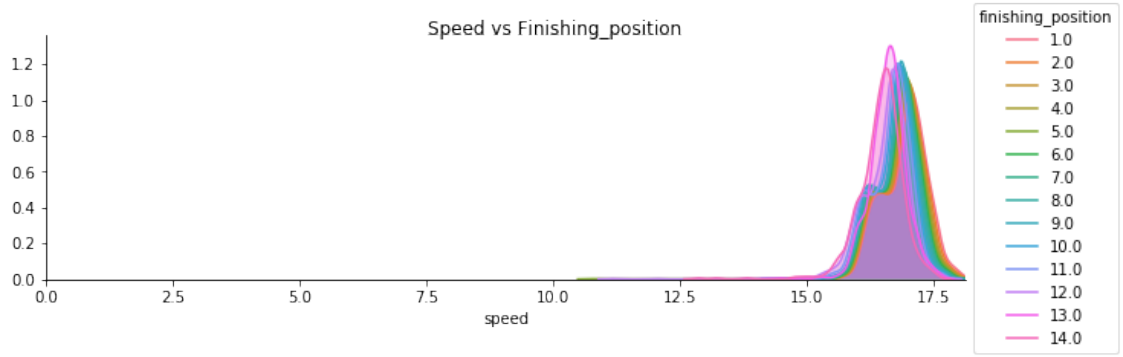
Winning Percentage of Jockey vs Finishing Position We can say that there is a realtionship between jockey performance and finishing position based on common sense. But we don't know the impact of a jockey on horse racing by guessing. However, somebody believes a good horse will win despite the jockey.

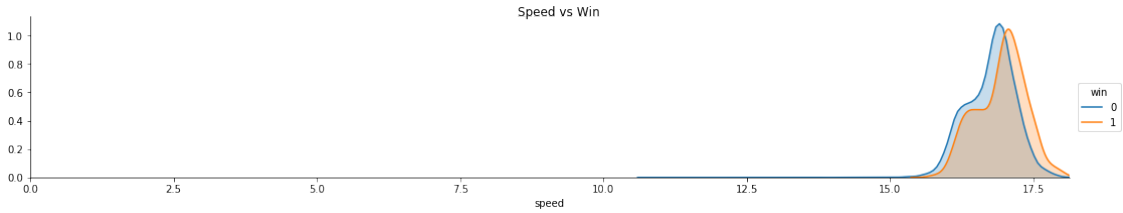




As we can see from the graphs above, it shows that the winning percentage of jockey is less than 10 have much higher rates of not winning races based on our data set. It is proved that there is a strong realtionship between jockey performance and finishing position.

Speed vs Finishing Position Speed should be one of the most important variables when we are picking the winner of a race. In the data wrangling, we divided the race distance(meter) by finish time(second) to get the speed. Therefore the speed is measured by meter per second.





As we can see from the graphs above, it shows that the speed faster than 17 meters per second has a better chance of winning. It is proved that there is a realtionship between speed and finishing position but not a strong relationship.

1.3 Logistic Regression

1.3.1 Converting Categorical Features

```
C:\Users\Jim\Anaconda3\lib\site-packages\pandas\core\frame.py:3697: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  errors=errors)
```

1.3.2 Train Test Set Split

```
Out[191]: 29364
```

1.3.3 Stat model

```
Optimization terminated successfully.
         Current function value: 0.239382
         Iterations 9
```

Out[257]:

```
<class 'statsmodels.iolib.summary2.Summary'>
"""
                           Results: Logit
=================================================================
Model:                Logit              Pseudo R-squared:   0.147
Dependent Variable:   win                AIC:                9866.5083
Date:                 2018-07-31 02:10   BIC:                9969.6088
No. Observations:     20554              Log-Likelihood:     -4920.3
Df Model:             12                 LL-Null:            -5768.0
Df Residuals:         20541              LLR p-value:        0.0000
Converged:            1.0000             Scale:              1.0000
No. Iterations:       9.0000
-----------------------------------------------------------------
                        Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
actual_weight         -0.0314   0.0044  -7.1739 0.0000 -0.0400 -0.0228
declared_horse_weight -0.0008   0.0005  -1.7911 0.0733 -0.0017  0.0001
draw                  -0.0114   0.0074  -1.5433 0.1228 -0.0259  0.0031
speed                  0.2099   0.0393   5.3406 0.0000  0.1329  0.2869
win_odds              -0.0691   0.0036 -18.9541 0.0000 -0.0763 -0.0620
race_distance         -0.0000   0.0001  -0.3006 0.7637 -0.0002  0.0002
jockey_1st             0.0480   0.0099   4.8652 0.0000  0.0287  0.0673
jockey_2nd            -0.0020   0.0177  -0.1134 0.9097 -0.0367  0.0327
jockey_3rd            -0.0174   0.0157  -1.1073 0.2681 -0.0481  0.0134
trainer_1st            0.0765   0.0150   5.0904 0.0000  0.0470  0.1059
trainer_2nd           -0.0512   0.0219  -2.3387 0.0194 -0.0941 -0.0083
trainer_3rd           -0.0669   0.0262  -2.5551 0.0106 -0.1182 -0.0156
daysince               0.0006   0.0009   0.7413 0.4585 -0.0010  0.0023
=================================================================

"""
```

1.3.4 logistic regression model

Out[194]:

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

Out[195]:

	Lose_Prob	Win_Prob
0	0.767215	0.232785
1	0.641847	0.358153

```
2   0.973122  0.026878
3   0.991069  0.008931
4   0.912457  0.087543
```

Out[199]:

	race_id	actual_weight	declared_horse_weight	draw	speed	win_odds	\
0	2016-185	125.0	1101.0	3.0	16.829704	7.4	
1	2016-472	115.0	1077.0	9.0	16.910936	3.1	
2	2016-506	126.0	1086.0	3.0	16.123831	30.0	
3	2016-612	121.0	1068.0	12.0	16.396803	46.0	
4	2016-472	115.0	1159.0	8.0	16.800448	11.0	

	race_distance	jockey_1st	jockey_2nd	jockey_3rd	...	Griffin Race	\
0	1600	24.691992	15.092402	12.782341	...	0	
1	1800	24.691992	15.092402	12.782341	...	0	
2	2000	11.842105	10.526316	13.157895	...	0	
3	2400	6.627566	8.680352	8.621701	...	0	
4	1800	7.085714	9.371429	6.971429	...	0	

	Group One	Group Three	Group Two	Hong Kong Group One	\
0	0	0	0	0	
1	0	0	0	0	
2	0	0	0	0	
3	0	1	0	0	
4	0	0	0	0	

	Hong Kong Group Three	Hong Kong Group Two	Restricted Race	win	Win_Prob
0	0	0	0	0	0.232785
1	0	0	0	1	0.358153
2	0	0	1	0	0.026878
3	0	0	0	0	0.008931
4	0	0	0	0	0.087543

[5 rows x 31 columns]

Out[242]:

	race_id	actual_weight	declared_horse_weight	draw	speed	\
3	2016-612	121.0	1068.0	12.0	16.396803	
128	2016-612	113.0	1119.0	11.0	16.174687	
266	2016-612	127.0	1074.0	4.0	16.497113	
305	2016-612	122.0	978.0	7.0	16.451878	
344	2016-612	124.0	1096.0	9.0	16.483516	
3971	2016-612	127.0	1269.0	8.0	16.314323	
6477	2016-612	115.0	1211.0	3.0	16.039564	
9991	2016-612	113.0	1006.0	5.0	16.371078	
10666	2016-612	120.0	1145.0	10.0	15.957447	
11578	2016-612	115.0	1105.0	1.0	16.140964	
13318	2016-612	130.0	1218.0	13.0	16.055660	
13896	2016-612	118.0	1022.0	6.0	16.441735	
14412	2016-612	133.0	1097.0	2.0	16.262366	

```
       win_odds  race_distance  jockey_1st  jockey_2nd  jockey_3rd  \
3          46.0           2400    6.627566    8.680352    8.621701
128        80.0           2400    3.734440    7.261411    6.016598
266         2.1           2400   24.691992   15.092402   12.782341
305         6.5           2400    9.221902   11.239193    9.221902
344         8.0           2400    9.630102    9.757653    9.757653
3971       99.0           2400    8.203678    9.335219    7.213579
6477       99.0           2400    5.263158    5.623648    6.560923
9991       44.0           2400    4.828974    4.828974    6.304494
10666      95.0           2400    6.019417    7.961165   11.650485
11578      29.0           2400    5.940594    6.435644    6.789250
13318      70.0           2400    8.163265    7.755102    6.326531
13896       8.1           2400    7.085714    9.371429    6.971429
14412       4.6           2400   17.442582   14.400993    9.310987

            ...       Group Three  Group Two  Hong Kong Group One  \
3           ...                 1          0                    0
128         ...                 1          0                    0
266         ...                 1          0                    0
305         ...                 1          0                    0
344         ...                 1          0                    0
3971        ...                 1          0                    0
6477        ...                 1          0                    0
9991        ...                 1          0                    0
10666       ...                 1          0                    0
11578       ...                 1          0                    0
13318       ...                 1          0                    0
13896       ...                 1          0                    0
14412       ...                 1          0                    0

       Hong Kong Group Three  Hong Kong Group Two  Restricted Race  win  \
3                          0                    0                0    0
128                        0                    0                0    0
266                        0                    0                0    1
305                        0                    0                0    0
344                        0                    0                0    0
3971                       0                    0                0    0
6477                       0                    0                0    0
9991                       0                    0                0    0
10666                      0                    0                0    0
11578                      0                    0                0    0
13318                      0                    0                0    0
13896                      0                    0                0    0
14412                      0                    0                0    0

       Win_Prob  Pred_Winner  Favor_Winner
3      0.008931            0             0
```

```
            128    0.000625            0             0
            266    0.302963            1             1
            305    0.128345            0             0
            344    0.136612            0             0
            3971   0.000148            0             0
            6477   0.000148            0             0
            9991   0.013701            0             0
            10666  0.000141            0             0
            11578  0.036689            0             0
            13318  0.001254            0             0
            13896  0.160046            0             0
            14412  0.190044            0             0

            [13 rows x 33 columns]
```

Out[240]: 177600.0

Out[245]: 1171

Out[241]: 95300.0

Out[218]: 0.28963153384747214