*Jordi Serra*
*Paula Sobrevals*

# PRÀCTICA2: Neteja i anàlisi de les dades

**Link al GitHub:** https://github.com/PSobrevals/PRAC2_NetejaAnalisiDades
**Document PDF amb les preguntes i respostes a respondre de la PRAC2.**

**Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:**

1. **Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?**

   Hem obtingut el dataset de la pàgina web suggerida: Kaggle (https://www.kaggle.com/lantanacamara/hong-kong-horse-racing?select=race-result-race.csv).

   El dataset és compost de dos taules, race-result-horse.csv i race-result-race.csv. Nosaltres hem ajuntat aquestes dues taules, utilitzant un merge() amb la columna de race_id que coincideix a totes dues taules, creant un únic dataset.

   Aquest conté els resultats de les 1561 curses de cavalls fetes a Hong Kong entre el 14 de Setembre del 2014 i el 16 de Juliol de 2017. Així com, informació sobre els cavalls i genets que hi van participar i els seus entrenaments, com informació sobre les condicions de les curses.

   Les carreres de cavalls a Hong Kong són un gran atractiu per a la població i el turisme, a més a més són de les més populars i es segueixen online des de molts llocs del món. Al ser tan populars, generen molta expectació i les apostes són bastant comuns. Poder endevinar el cavall guanyador és un repte que molta gent segueix intentant i apostant-hi.

   En aquesta pràctica pretenem crear un model per a predir els resultats de les carreres de cavalls, és a dir, pretenem respondre la següent pregunta: Qui serà el cavall guanyador a la següent carrera?

2. **Integració i selecció de les dades d'interès a analitzar.**

   Un cop tenim el dataset sencer, és a dir, un cop hem ajuntat les dues taules per a format una sola amb tota la informació, visualitzem de quants elements és formada: 30189 observacions i 30 variables.

Hem fet un cop d'ull a les variables i hem decidit que algunes no calia usar-les, ja que algunes eren redundants, estaven repetides o no aportaven una informació vàlida:

1. "incident_report": Són comentaris sobre la cursa, no utilitzarem aquesta variable per als nostres anàlisis, ja que no ens aporta informació necessària.

2. "running_position_X": Són les posicions intermitges de la carrera. Aquesta informació és repetida, ja que ja tenim els resultats finals en la columna "finishing_position".

3. "SRC": Tampoc ens aporta res més que el nom dels fitxers de dades.

4. "horse_number": Depèn de cada cursa i no identifica el cavall, per tant deixem només el "horse_name".

5. "race_number": També és redundant amb les altres variables, tenim el "race_id".

6. "race_date": La data de la carrera és indiferent per l'anàlisi de les dades.

7. "sectional_time": Tindrem en compte només el temps total de la carrera, no de cada volta.

8. "horse_id" : Ja tenim un identificador pel cavall, aquest és redundant.

9. "length_behind_winner" : És redundant, ja tenim la posició final de la cursa.

Per tant, de moment utilitzarem les següents columnes per a fer els nostres anàlisis:

| | | | |
|---|---|---|---|
| race_id | finishing_position | horse_name | jockey |
| draw | declared_horse_weight | track | finish_time |
| race_class | race_distance | track_condition | race_name |
| trainer | actual_weight | win_odds | race_course |

## 3. Neteja de les dades.

### 3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Primer, visualitzem el dataset i ens n'adonem que no tots els elements buis estan representats de la mateixa manera: Alguns estan representats com a "---", "N","SH", o "HD". Per a poder treballar millor, hem de tenir tots els missing values amb la mateixa representació, així que els hem modificat per a tenir la notació "NA". Un cop ja tenim tots els valors buits amb la mateixa identificació, podem mirar si tenim gaires elements buits, i en quines variables:

```
      race_id    finishing_position           horse_name                jockey
            0                   825                    0                    29
      trainer         actual_weight declared_horse_weight                  draw
            0                     0                    0                   591
  finish_time              win_odds          race_course            race_class
          669                   591                    0                     0
race_distance       track_condition             race_name                 track
            0                     0                    0                     0
```

Podem observar que tenim alguns valors buits al nostre dataset, primer ens centrarem en la variable "finishing_position". Determinem quin % de valors buits representa, per a poder decidir com actuar en conseqüència:

```
Proporció de valors nulls de la variable finishing_position: 2.73 %
```

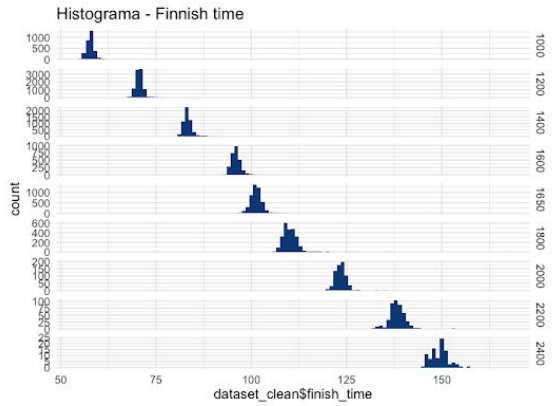
Veiem que representa un valor molt baix de totes les nostres dades, així que aquest cas el podem gestionar eliminant les observacions nules. Si representés una proporció més gran de les nostres dades, al voltant de 15-20% hauriem d'imputar els missing values, per exemple utilitzant un anàlisi de regressió.

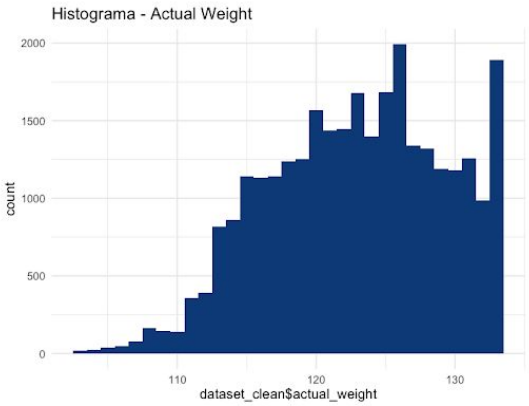
Un cop eliminades les observacions nules, veiem que tots els altres valors nuls han quedat conseqüentment eliminats també. Ja no tenim cap missing data.

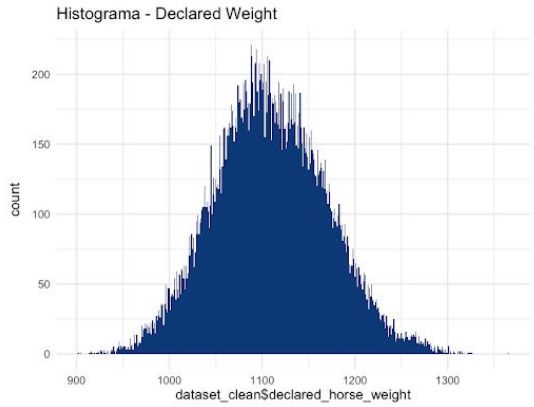
### 3.2. Identificació i tractament de valors extrems.

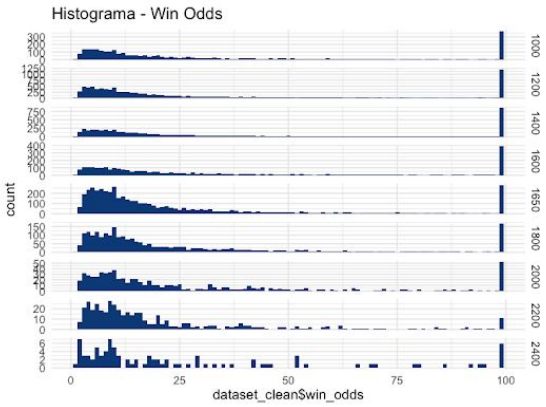
Es fa un estudi de les variables numèriques continues, ja que les categòriques no tenen valors extrems. Per a fer-ho utilitzarem una representació gràfica (histograma i boxplot) que ens permetrà identificar i visualitzar més fàcilment els

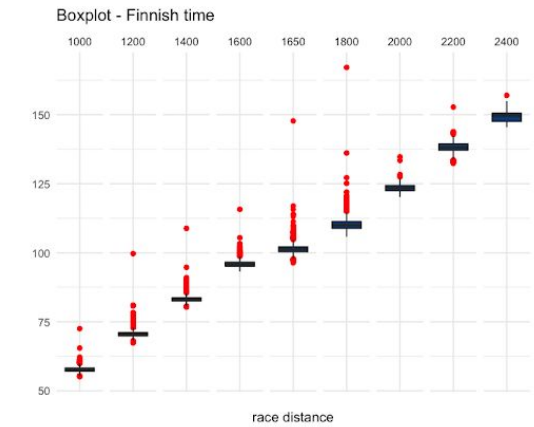
valors extrems. El nostre dataset conté curses de diferents llargades, aquest fet, altera els valors d'algunes variables (per exemple, el temps d'una cursa de 1000m serà molt diferent al d'una cursa de 2400m) per tant, per a analitzar-les cal separar els valors segons la llargada de la cursa:

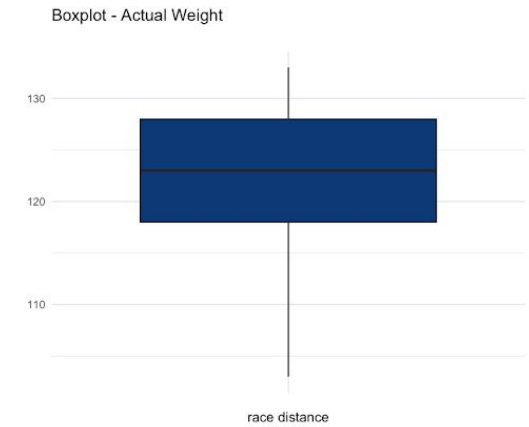


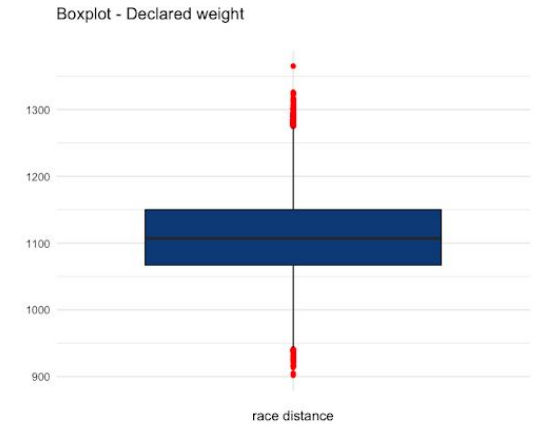



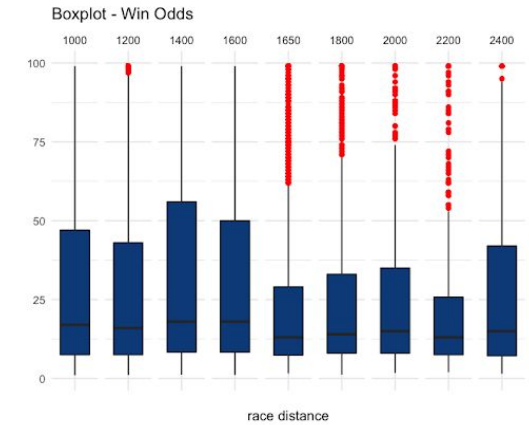



Les gràfiques obtingudes a partir de l'histograma i dels boxplot, mostren, per a cada una de les diferents llargades de cursa, els diferents valors extrems de les variables numèriques estudiades (en vermell).

Per la variable del temps de finalització de cada carrera (Finish time), podem veure que hi ha cavalls que triguem especialment molt més que la resta en acabar la cursa. Podem veure també, com el pes del genet (Actual weight) està esbiaixat a l'alça, però tot i així hi ha alguns valors baixos, però no estan considerats outliers. El mateix passa amb el pes dels cavalls (Declared weight), que la majoria estan distribuïts centralment, però tot i això hi ha valors que queden marcats com a extrems. Podem veure que la variació de pes és molt petita i sembla que els cavalls més grans son els que fan curses més curtes, però no és significatiu.

Eliminarem tots aquells valors outliers. Tot i així, podem observar que a la variable "win_odds", tots els outliers són els cavalls que tenen més probabilitats de guanyar. Com que el nostre estudi té a veure amb aquests, no eliminarem aquests outliers del nostre dataset i els tindrem en compte pels anàlisis posteriors.

```
"Hem eliminat 6346 outliers."
"Ara el nostre dataset té 23018 observacions."
```

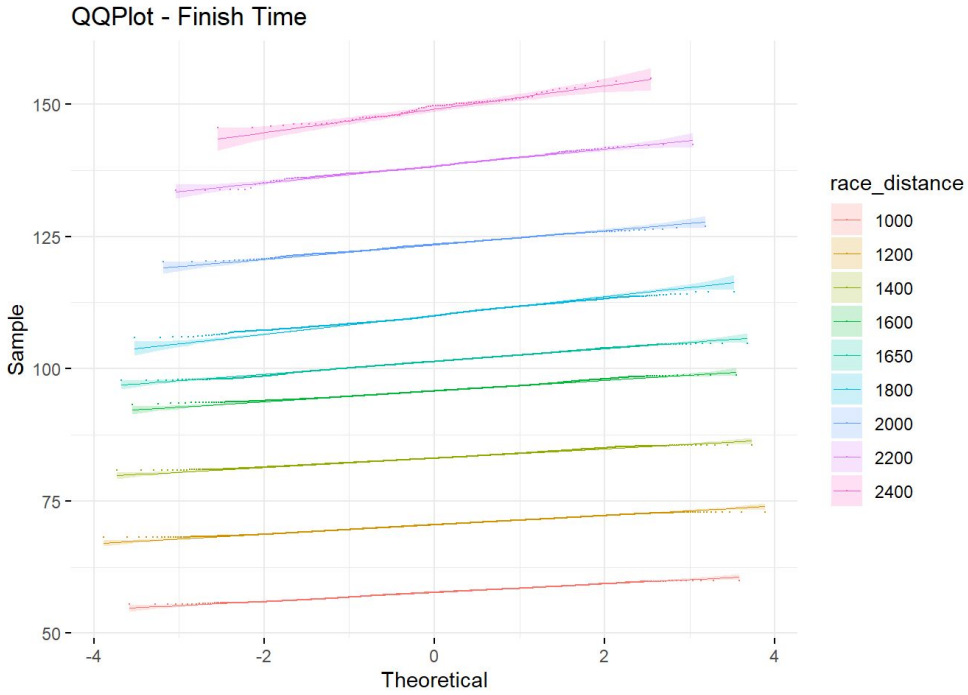
## 4. Anàlisi de les dades.

### 4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Anem a descriure com hem analitzat les dades que té el dataset. Com s'ha dit hi ha dades que són númeriques i d'altres que són categoriques, hi ha que aporten valor a l'estudi i altres que no, ja que són completament redundats a l'hora de fer l'estudi, com poden ser l'identificador del cavall, o el número. Aquestes dades ja s'han filtrat en el primer pas, en la neteja.

La variable depenent que hem seleccionat per aquest estudi és la posició final del cavall en cada una de les curses, tindrem en compte per poder trobar aquesta posició final totes les altres variables que hem fet servir després de netejar el dataset. En l'estudi de la regressió de les variables ens hem adonat que a mesura que en el càlcul de d'aquesta anem afegint les variables el resultat obtingut millora sempre, ja que no s'ha mostrat cap correlació entre les dades del dataset netejat, el dataset_clean. Així doncs, la inclusió del grup de dades independents fa que millori el resultat en la predicció del valor depenent.

El que si cal notar és que com ja descrit, la distancia de la cursa si afecta clarament a la resta de variables, sobretot al temps de finalització i a la posició, per tant l'estudi que fem de les dades a partir d'aquí es fa per cada una de les curses per separat, per a que no afecti els diferents valors d'aquesta variable. Podem veure com estan distribuïdes aquestes dades si mirem la següent gràfica obtinguda en l'anàlisis.




Es poden veure com estan diferenciades les diferents carreres.

## 4.2. Comprovació de la normalitat i homogeneïtat de la variància.

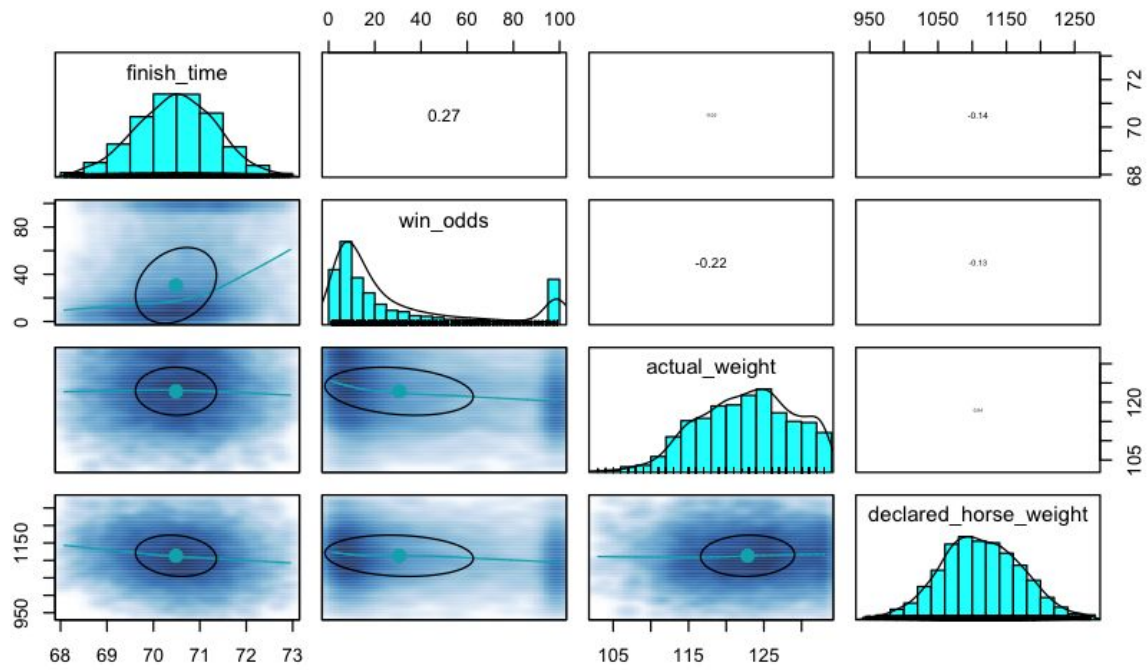
Per a comprovar la normalitat i la homogeneïtat de la variança, hem de mirar com estan distribuïdes les dades, per a comprovar si compleixen amb la hipòtesi de la distribució normal de les dades, i si es comporten de manera homogènia. Ho farem de dues maneres:

a) De manera visual: Per a comprovar la normalitat utilitzarem un QQPlot, que ens permetrà veure com es comporten les dades i si estan ajustades a la distribució normal. Per a comprovar la homogeneitat utilitzarem un scatter plot que ens ajudarà a veure si es comporten de manera homogènia.

b) Utiltzant tests específics: Per quantificar la normalitat usarem el Test de Shapiro–Wilk, el qual és considerat un dels test més potents per a contrastar la normalitat. Per quantificar la homogeneitat, usarem el Lavene Test i el Fligner-Killeen test segons si les nostres dades segueixen una distribució normal o no.

Un cop hem generat els plots, hem vist que la majoria de les variables segueixen una distribució normal, excepte la variable “winn_odds”, que sembla aproximar-se a una distribució Bernoulli.

En la següent figura, podem visualitzar les homogeneitats entra variables (a la meitat baixa de la diagonal) i la distribució que aquestes presenten (als histogrames).




En el plot, veiem que quasi no hi ha homogeneïtat entre les variables, tot i així algunes variables presenten petites homogeneïtats: “win_odds” presenta una homogeneitat positiva amb amb “finish_time”, i una petita homogeneitat negativa amb “actual_weight”. “Finish_time” també presenta una petita homogeneïtat negativa amb “declared_horse_weight”.

Un cop hem fet els tests estadiístics, Test de Shapiro–Wilk per a quantificar la normalitat i Lavene Test i el Fligner-Killeen test per a quantificar la homogeneïtat, hem vist que tots ells eren significants:

| | **Normalitat** | **Homogeneïtat** |
|---|---|---|
| **Finish time** | Shapiro-Wilk normality test<br>W = 0.99754, p-value < 1.1e-16 | Levene's Test for Homogeneity of Variance (center = median)<br>Df F value Pr(>F)<br>group 8 366.78 < 2.2e-16 ***<br>28078 |
| **Actual weight** | Shapiro-Wilk normality test<br>W = 0.97421, p-value < 2.2e-16 | Levene's Test for Homogeneity of Variance (center = median)<br>Df F value Pr(>F)<br>group 8 18.558 < 2.2e-16 ***<br>28078 |
| **Declared horse weight** | Shapiro-Wilk normality test<br>W = 0.99683, p-value = 8.31e-09 | Levene's Test for Homogeneity of Variance (center = median)<br>Df F value Pr(>F)<br>group 8 2.2772 0.01968 *<br>28078 |
| **Win odds** | Shapiro-Wilk normality test<br>W = 0.76156, p-value < 2.2e-16 | Fligner-Killeen test of homogeneity of variances<br>med chi-squared = 401.25,<br>df = 8, p-value < 2.2e-16 |

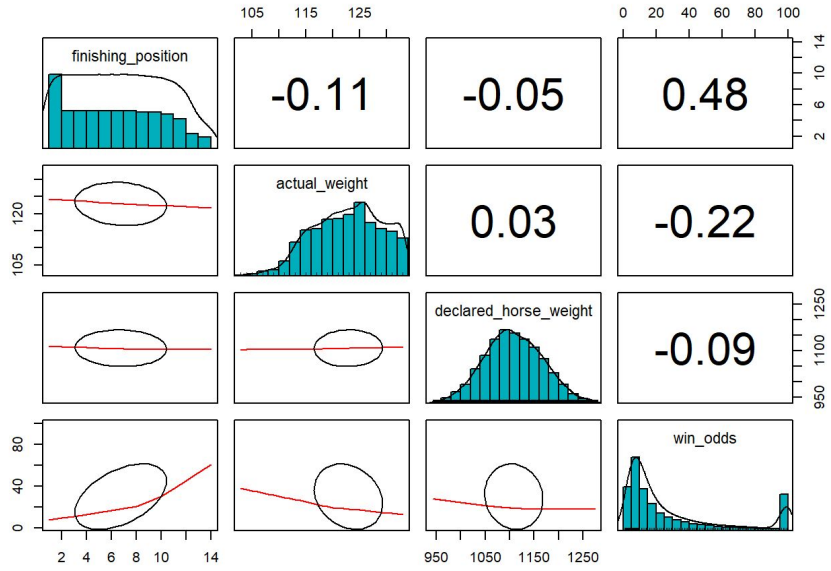
**4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis,correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.**

Correlació:

```
##                       finishing_position actual_weight declared_horse_weight
## finishing_position                 1.00         -0.11                 -0.05
## actual_weight                     -0.11          1.00                  0.03
## declared_horse_weight             -0.05          0.03                  1.00
## win_odds                           0.48         -0.22                 -0.09
##                       win_odds
## finishing_position        0.48
## actual_weight            -0.22
## declared_horse_weight    -0.09
## win_odds                  1.00
##
## n= 28087
##
##
## P
##                       finishing_position actual_weight declared_horse_weight
## finishing_position                     0             0
## actual_weight          0                             0
## declared_horse_weight  0                 0
## win_odds               0                 0           0
##                       win_odds
## finishing_position     0
## actual_weight          0
## declared_horse_weight  0
## win_odds
```

En aquest cas, hem fet servir tres tècniques estadístiques a les dades, al grup de dades dependents que són numerics hem aplicat la correlació directament, fent servir els algoritmes de PEARSON. Aquest estudi ens ha demostrar que aquestes variables no estan correlacionades entre sí, tal i com podem veure als resultats obtinguts.

O mitjançant la imatge següent,




En el cas de la resta de variables categòriques que també es fan servir com a variables independents hem aplicat la funció ANOVA factoritzant aquestes variables per tal de poder efectuar l'estudi estadístic. Podem veure en els resultats obtinguts que la funció anova ens dóna un p-value significant, menor de 0'05, per totes les variables, excepte en el cas d'una, la variable *track_condition* que es comporta diferent de totes les altres.

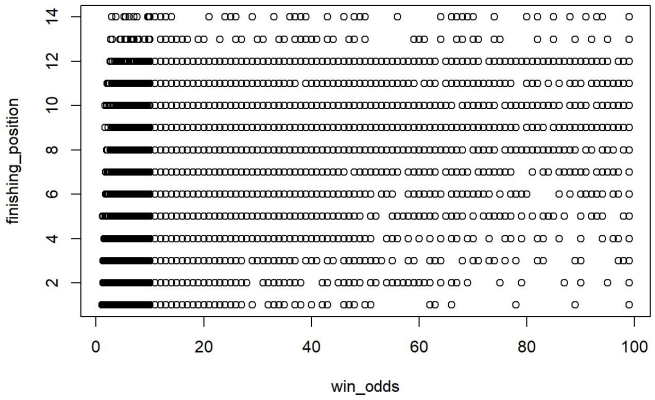
```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## horse_name      2125  85123    40.1   4.021  < 2e-16 ***
## jockey            83  16081   193.7  19.450  < 2e-16 ***
## trainer           23    925    40.2   4.038 2.44e-10 ***
## draw              14   5597   399.8  40.138  < 2e-16 ***
## race_course        1   2162  2161.5 216.994  < 2e-16 ***
## race_class        15  11219   747.9  75.083  < 2e-16 ***
## race_distance      8    572    71.5   7.179 1.52e-09 ***
## track_condition    8    123    15.4   1.542    0.137
## track              6    532    88.7   8.908 9.75e-10 ***
## Residuals      25803 257030    10.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Jordi Serra*
*Paula Sobrevals*

Regresió:

Un cop hem demostrat que no hi ha correlació directe entre les variables independents anem a mirar la regressió entre aquestes variables.

Per fer això mirem primer directament la relació que pot haver entre les variables, respecte a la variable dependent que tenim, la posició final. Ho podem fer com a primera aproximació mirant directament les dades. En aquest cas, mirant totes les gràfiques de les dades es pot veure que es comporten aproximadament igual. Es mostra una gràfica de la distribució de les dades a tall d'exemple entre la variable dependent i la probabilitat de guanyar a priori que té el cavall, abans de començar la cursa, és a dir, la variable *win_odds*.




Es pot veure com la gran majoria de dades estan totes relacionades amb la posició final que obté a partir de la que s'havia pensat amb les probabilitats inicials.

També ho podem fer mitjançant la comanda *lm* de R per tant que ens mostri la relació que hi ha entre les diferents variables que estem relacionant a la comanda.

Si tenim en compte totes les variables independents a l'hora de calcular la regressió directe amb la variable dependent, veiem que, com s'ha comentat abans, a mesura que incloem més variables els resultats obtinguts són més bons.

```
##
## Call:
## lm(formula = finishing_position ~ actual_weight + declared_horse_weight +
##     win_odds + trainer + jockey + horse_name + race_class + race_course +
##     track_condition + race_name + race_id, data = dades)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.851 -1.796  0.000  1.693 10.144
##
## Coefficients: (593 not defined because of singularities)
##                                                                          Estimate
## (Intercept)                                                             -5.521307
## actual_weight                                                            0.092589
## declared_horse_weight                                                   -0.003926
## win_odds                                                                 0.026669
## trainerA S Cruz                                                         -0.072254
```

...

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.985 on 7141 degrees of freedom
## Multiple R-squared:  0.4724, Adjusted R-squared:  0.2884
## F-statistic: 2.567 on 2491 and 7141 DF,  p-value: < 2.2e-16
```

No acabem obtenint un valor de 1 o molt molt aproximat, ja que se la posició en la carrera de cavalls dependrà de molts més factors dels que tenim en el dataset, com per exemple podria ser l'edat del cavall, el descans, les hores d'entrenament, tant pròpies com dels altres cavalls.

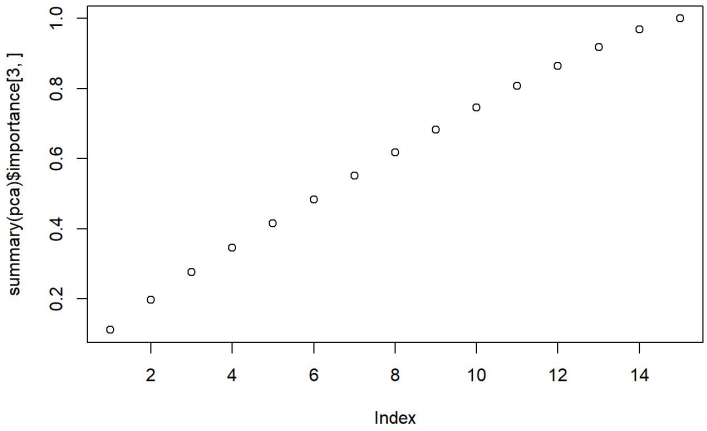
PCA+Regressió

Aquesta darrera prova la farem per a comparar-la amb la regressió estàndard. Volem veure si podem millorar el model creat anteriorment amb la regressió, utilitzant els principal components necessaris com a predictors per a ajustar el model de manera més precisa. Un cop s'ha executat la comanda tenim els components principals que queden de la següent manera.

```
Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.16070 1.11797 1.08560 1.02418 1.01674 1.00906 1.00575
## Proportion of Variance 0.09623 0.08927 0.08418 0.07492 0.07384 0.07273 0.07225
## Cumulative Proportion  0.09623 0.18551 0.26969 0.34461 0.41845 0.49118 0.56343
##                            PC8     PC9    PC10    PC11    PC12    PC13   PC14
## Standard deviation     0.99864 0.98577 0.96839 0.95630 0.90073 0.8759 0.84387
## Proportion of Variance 0.07123 0.06941 0.06698 0.06532 0.05795 0.0548 0.05087
## Cumulative Proportion  0.63466 0.70407 0.77106 0.83638 0.89433 0.9491 1.00000
```

La proporció de variança a cumulada pels diferents components podem veure que és molt baixa. Necessitem 7 PC per a poder explicar el 50% de la variança. Això ens comença a donar una idea que els PC no seran molt bona opció per a millorar el model. Tot i així, ho comprovarem.

En format d'imatge podem veure els components principals de la mateixa manera, la següent imatge mostra tots els components principals.




Un cop tenim la PCA feta, hem generat una regressió com l'anterior:

```
Call:
lm(formula = finishing_position_y ~ ., data = dataset_pcs)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58189 -0.64952 -0.04013  0.62297  2.94892

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.258e-16  5.089e-03   0.000 1.000000
PC1          2.640e-01  4.385e-03  60.213  < 2e-16 ***
PC2          2.069e-01  4.552e-03  45.457  < 2e-16 ***
PC3         -1.020e-01  4.688e-03 -21.767  < 2e-16 ***
PC4         -5.479e-02  4.969e-03 -11.026  < 2e-16 ***
PC5         -7.691e-05  5.006e-03  -0.015 0.987741
PC6         -4.229e-02  5.044e-03  -8.385  < 2e-16 ***
PC7          5.494e-02  5.060e-03  10.857  < 2e-16 ***
PC8          1.700e-02  5.096e-03   3.336 0.000851 ***
PC9         -3.281e-02  5.163e-03  -6.355 2.12e-10 ***
PC10        -8.349e-02  5.256e-03 -15.885  < 2e-16 ***
PC11         6.212e-02  5.322e-03  11.672  < 2e-16 ***
PC12        -1.999e-02  5.650e-03  -3.539 0.000403 ***
PC13        -2.405e-02  5.810e-03  -4.140 3.49e-05 ***
PC14        -3.613e-01  6.031e-03 -59.905  < 2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8529 on 28072 degrees of freedom
Multiple R-squared:  0.2729,     Adjusted R-squared:  0.2725
F-statistic: 752.4 on 14 and 28072 DF,  p-value: < 2.2e-16
```

Veiem que tal i com hem predit, els resultats de la regressió amb el PCA han empitjorat en relació amb l'anterior. Veiem que aquest model té un R2 = 0.27, aquest valor és molt menys que l'anterior (R2 = 0.47). És a dir, que utilitzar els PCA per a millorar el model no ha donat els resultats esperats i no l'hem pogut millorar.

## 5. Representació dels resultats a partir de taules i gràfiques.

Hem decidit utilitzar un gràfic interactiu, generat amb Plotly, on podem decidir el cavall, i la llargada de cursa a representar. El gràfic és ordenat cronològicament, i en l'eix d'ordenades hi trobem el temps final de la cursa. A més a més, els punts genrats estan colorejats segons la llargada de la cursa, i tenen la mida segons el resultat final (aquells cavalls guanyadors tindran una mida més petita, que els perdedors). Corrent el ratolí per sobre indica la informació del punt en qüestió.

Aquí tenim un exemple del gràfic:

Per a respondre la resta de preguntes ens hem anat recolzant en diferents gràfics i taules, que podem veure distribuïdes al llarg del pdf.

6. **Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?**

   Inicialment ens havíem marcat com a objectiu a partir del dataset escollit, crear un de sol i poder trobar la manera de trobar per a qualsevol cavall la posició en que acabaria en una de les curses. És a dir, que a partir de les variables que hi ha al dataset, exceptuant clar el temps de finalització, poder inferir el resultat amb la resta de variables.

   Un cop netejades les dades, treballant les diferents curses per separat, ja que molts cavalls només participen en alguns tipus de curses, poques, i no en totes, hem vist que la variable dependent del dataset, la posició final, depèn de totes i cada una de les diferents variables independents, però tot i això, degut a la complexitat que resulta el problema que ens hem plantejat, es veu com amb les variables obtingudes del dataset no son suficients per obtenir una regressió molt significativa amb la variable que volem obtenir a partir de la resta. Amb les que tenim hem obtingut una regressió del 0'5. Les variables incloses en el dataset de les curses de cavalls no són suficients per obtenir una regressió molt més propera a 1. Amb altres variables com pot ser el temps d'entrenament del cavall, el temps de descans abans de les curses, l'alimentació o altres factors que poden influir en l'estat de l'animal.

7. **Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.**

   El codi complert i funcional és en RMarkdown i es troba al GitHub amb el nom: *practica2.Rmd*. També hi és en format html: *practica2.html*

   https://github.com/PSobrevals/PRAC2_NetejaAnalisiDades

*Jordi Serra*
*Paula Sobrevals*

**Taula de contribucions al treball:**

| **Contribucions** | **Signatura** | |
|---|---|---|
| | **Jordi** | **Paula** |
| Recerca Prèvia | ✔ | ✔ |
| Redacció de les respostes | ✔ | ✔ |
| Desenvolupament codi | ✔ | ✔ |