

# GRUPPO: CORES

## STATISTICA COMPUTAZIONALE – REPORT FINALE

Maccianti Federico   Rapacioli Nicola   Riva Pietro  
(909656)   (915439)   (908813)

## 1 Introduzione

Il seguente studio, si pone come obiettivo quello di individuare i diversi stili di guida presenti nella Formula 1, tramite l'analisi dei dati di telemetria ([TracingInsights](#)<sup>1</sup>) di ciascun pilota. Lo studio si concentra sui dati relativi alle sessioni di qualifica della stagione di Formula 1 2025, selezionando per ciascun pilota il singolo giro migliore.

Il dataset originale include le seguenti variabili:

Variabile	Unità di misura	Supporto	Tipo
Gran premio	–	{Nomi GP}	character
Pilota	–	{Sigle Piloti}	character
Tempo dal via	s	$\mathbb{R}^+$	numeric
Distanza percorsa	m	$\mathbb{R}^+$	numeric
Distanza relativa	–	$[0, 1]$	numeric
Velocità	km/h	$\mathbb{N}$	numeric
Regime motore	RPM	$\mathbb{N}$	numeric
Marcia	–	$\{1, \dots, 8\}$	integer
Freno	–	$\{0, 1\}$	factor
Acceleratore	%	$[0, 100]$	numeric
DRS	–	$\{0, 1\}$	factor
Accelerazione laterale e longitudinale	g	$\mathbb{R}$	numeric
Coordinate spaziali x,y,z	m	$\mathbb{R}$	numeric
Tempo Giro	s	$\mathbb{R}^+$	numeric

Il dataset presenta le seguenti codifiche: le feature binarie assumono valore 0 in assenza dell'evento e 1 in sua presenza. La telemetria dell'acceleratore misura l'intensità dell'input del pilota, mentre per le distanze indicano la posizione progressiva rispetto allo start.

## 2 Analisi Esplorativa

### 2.1 Considerazioni sulle variabili

Poiché lo stile di guida non è riconducibile a variabili di tipo posizionale, le coordinate spaziali e le misure di distanza, sia assolute sia relative, vengono escluse dall'analisi.

In questa fase preliminare, lo stile di guida viene descritto attraverso variabili dinamiche, quali l'utilizzo dell'acceleratore e del freno e le accelerazioni longitudinali e laterali, che consentono, per esempio, di valutare rispettivamente le modalità di decelerazione in ingresso curva, l'intensità con cui la curva viene affrontata tutte caratteristiche che fanno riferimento allo stile di guida. A sostegno delle ipotesi sopra citate, si riporta la Figura 1 che confronta nel Gran Premio degli USA le accelerazioni per i piloti: Charles Leclerc, Lando Norris e Max Verstappen.

<sup>1</sup>Codice di estrazione dati nell'appendice A.

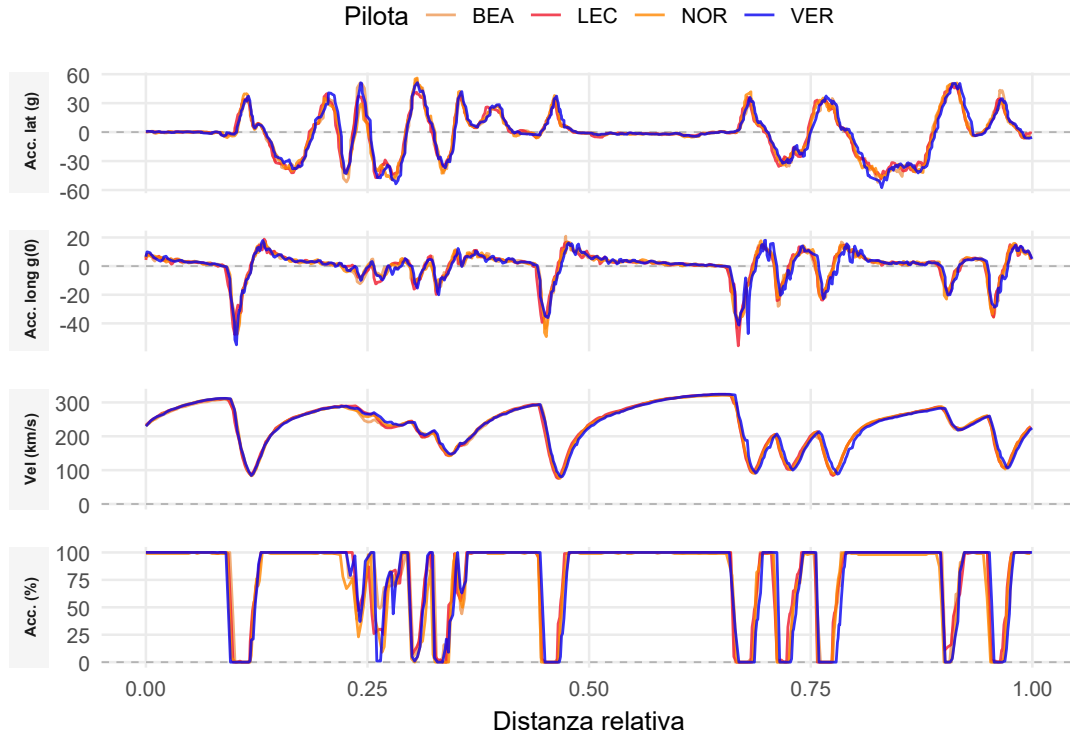


Figura 1: Accelerazioni VER, LEC, NOR.

Si nota infatti come nelle variazioni repentine si possano già individuare differenze significative tra piloti.

Dalle prime analisi descrittive sul dataset<sup>2</sup>, viene osservato come per il pilota Russell al Gran Premio di Miami siano presenti degli **NA** nella variabile **distanza relativa**. Osservando i dati grezzi si può notare come ciò sia riconducibile al malfunzionamento dei sensori telemetrici - presenza di molti zeri in molte variabili e accelerazioni costanti durante l'intero giro. Per i motivi elencati sopra, si procede dunque ad eliminare i record.

Infine, osservando la variabile **Tempo Giro**, alcuni piloti riportano alcuni valori **NA**. Osservando i dati<sup>2</sup>, ciò si verifica per i piloti Tsunoda, Bearman e Hadjar rispettivamente nei Gran premi di Emilia Romagna, Australia e Stati Uniti.

Effettuando un'analisi di tipo qualitativo, si scopre che ciò è riconducibile ad un incidente durante il giro di qualifica. Le uniche telemetrie disponibili sono infatti riconducibili al giro di riscaldamento delle gomme, come si nota nella figura 2. Al fine di eliminare rumore nell'analisi si rimuovono queste osservazioni.

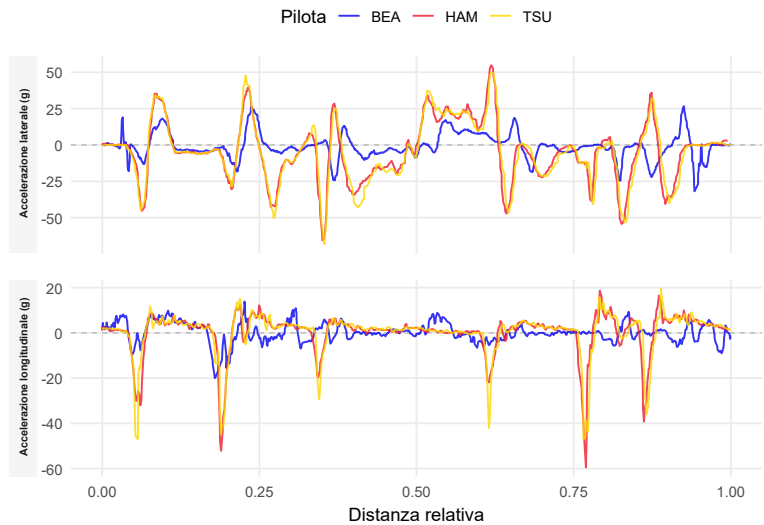


Figura 2: Confronto Australia Bearman

<sup>2</sup>Estratto codice R consultabile nell'appendice B.

Sempre per quanto riguarda la variabile **tempo giro**, considerando il regolamento della **FIA** (39.4.b.i ; PAG:47), secondo il quale se un pilota supera il 107% del tempo giro minore viene escluso dalle qualifiche, si uniforma il dataset. Incriminati risultano essere i piloti Hamilton, Tsunoda, Antonelli, Albon e Bortoletto nel Las Vegas Grand Prix e Stroll nel Dutch Grand Prix.

Analizzando il contesto, si nota che a Las Vegas i tempi più elevati sono imputabili alle condizioni meteo avverse (pioggia intensa), seguite da un progressivo asciugamento della pista che ha avvantaggiato gli altri piloti. Nel caso di Stroll, invece, il dato si riferisce a al primo giro di lancio completo, poiché nel momento del giro di qualifica non è stato completato il giro per un incidente.

Di conseguenza per eliminare il rumore, si elimina soltanto l'occorrenza di Stroll, poiché il giro non rappresenta il "migliore" ma solamente il quello di lancio.

### 2.1.1 Trasformazione e creazione di nuove variabili

Per quanto riguarda l'analisi della variabile di **accelerazione laterale**, viene trasformata tramite valore assoluto. Così da ottenere una misura dell'intensità complessiva (magnitudo)<sup>3</sup>.

Inoltre, l'**accelerazione longitudinale** viene suddivisa<sup>3</sup> in:

- **Accelerazione**, che comprende tutti i valori positivi;
- **Decelerazione** che al contrario della precedente comprende, tutti i valori negativi;

Successivamente viene applicato il valore assoluto e per le restanti variabili telemetriche si mantiene invece la forma originaria.

Si creano due nuove variabili di variazione percentuali<sup>4</sup>, per catturare l'intensità di variazione delle variabili che aiutano a distinguere lo stile di guida dei piloti, nei vari istanti di tempo:

- **\_lag1** : variazione percentuale rispetto all'osservazione precedente;
- **\_lag5** : variazione percentuale rispetto a cinque osservazioni precedenti;

che vengono calcolate come:

$$\Delta = \frac{x_t - x_{t-k}}{x_{t-k}}$$

indicando con  $x_t$  l'unità statistica al tempo  $t$  e  $k = \{1, 5\}$  numero di lag, per le misure di accelerazione e decelerazione. Nel caso in cui  $x_{t-k} = 0$  e  $x_t = 0$  non si applica la formula e la variabile **\_lag** assume 0 di default, quando invece si osserva solamente  $x_{t-k} = 0$  si sostituisce con  $x_{t-k} = 0.01$  per riuscire a mantenere la validità del calcolo e continuità delle dinamiche telemetriche.

## 2.2 Statistiche riassuntive

Vengono calcolate delle statistiche descrittive<sup>5</sup> al fine di valutare in che modo le variabili possano essere associate allo stile di guida, considerando separatamente ciascun Gran Premio e Pilota.

Per le misure di accelerazione laterale, longitudinale e decelerazione vengono calcolate media e deviazione standard.

La variabile relativa alla frenata non viene invece inclusa nell'analisi descrittiva, in quanto la sua natura binaria e la forte dipendenza dalle caratteristiche specifiche del singolo Gran Premio rendono poco informative misure sintetiche come media e deviazione standard. Tali indicatori risulterebbero infatti più rappresentativi del tracciato e delle condizioni di gara che dello stile di guida del pilota viene comunque presa in considerazione in modo latente dalla decelerazione.

Per le variabili di tipo **\_lagk** vengono calcolate medie e deviazioni standard distinguendo tra variazioni positive e negative, così da evidenziare eventuali asimmetrie nel comportamento dinamico del pilota.

<sup>3</sup>Estratto codice R consultabile nell'appendice C.

<sup>4</sup>Estratto codice R consultabile nell'appendice D

<sup>5</sup>Estratto codice R consultabile nell'appendice E.

Una volta calcolate media e deviazione standard, è possibile combinare questi valori per ottenere il coefficiente di variazione (CV):

$$CV = \frac{sd(x)}{mean(x)}$$

che misura la variabilità relativa di una variabile rispetto alla sua media. L'impiego di tale coefficiente permette di ridurre l'influenza del setup della vettura, rendendo confrontabili tra loro variabili che, in valore assoluto, dipendono dalle regolazioni meccaniche e aerodinamiche.

Si arriva così ad ottenere un dataset contenente 22 variabili, che forniscono informazione sullo stile di guida del pilota nella specifica gara.

Infine, per rimuovere l'influenza del tracciato, si riscalano tutte le variabili nell'intervallo  $[0, 1]$ .

Si procede dunque con una analisi delle correlazioni per ridurre la dimensionalità, e comprendere quali siano le variabili più significative nel fornire l'informazione.

## 2.3 Analisi delle componenti principali

Per limitare i casi di multicollinearità ed eliminare il rumore, avendo correlazioni molto elevate ( $cor(x, y) > 0.9$ )<sup>6</sup>, si cerca di eliminare le variabili rindondanti. Questo processo di scrematura permette di ridurre la dimensionalità a 20 variabili.

Seppur la dimensionalità sia diminuita, 20 dimensioni sono eccessive sia ai fini di interpretabilità che per peso computazionale per un Model Base Clustering.

Si effettua dunque un'analisi delle componenti principali (PCA) ai fini di ridurre la dimensionalità. I risultati ottenuti tramite PCA<sup>7</sup> mostrano come le prime 5 componenti spiegano più del 70% della varianza, come si nota nella figura 3.

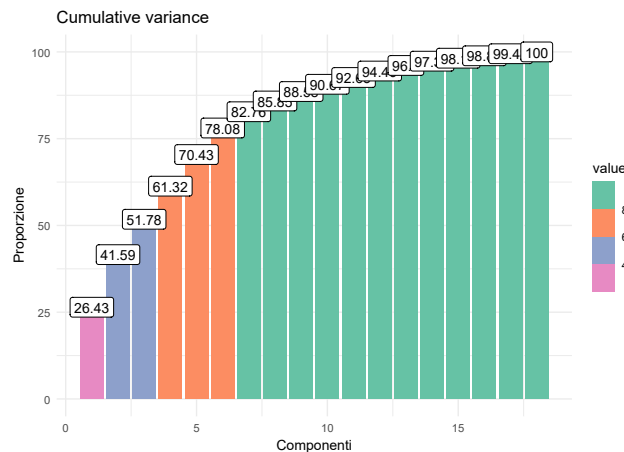


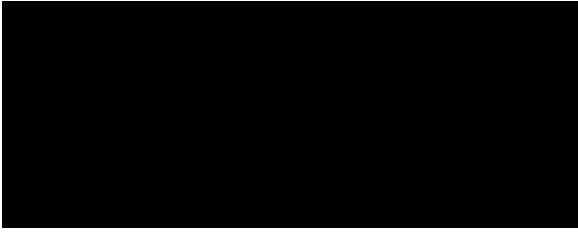
Figura 3: Varianza Spiegata Cumulata

### 2.3.1 Interpretazione delle componenti

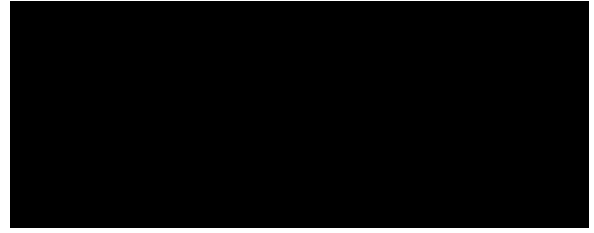
<sup>6</sup>Estratto del codice R consultabile nell'appendice F

<sup>7</sup>Estratto del codice R consultabile nell'appendice G

Come si evince dalla Figura 4, la distribuzione ...



(a) Distribuzione Variabile X



(b) Scatterplot X vs Y

Figura 4: Analisi esplorativa iniziale delle variabili principali.

### 3 Modellazione

Abbiamo applicato un modello di regressione lineare:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i \quad (1)$$

## A Codice R Commentato Estrazione dei dati

```
1 library(jsonlite)
2 library(tidyverse)
3 rm(list=ls())
4 #Cartella origine
5 cartella <- "C:/Users/feder/Documents/datasets/Computazionale/2025"
6
7 #filtro sui file con i giri e poi qualififiche
8 file_giri <- list.files(path = cartella,
9 pattern = "laptimes\\.json$",
10 full.names = TRUE,
11 recursive = TRUE)
12
13 files_qualififiche_lap <- file_giri[grepl("Qualifying", file_giri) & !grepl("Sprint", file_giri)]
14 lista_dati <- list()
15
16 #Ciclo di estrazione
17
18 for(i in 1:length(files_qualififiche_lap)) {
19
20   cartella_giro <- files_qualififiche_lap[i]
21
22   laptimes_data <- fromJSON(cartella_giro)
23
24   giro <- as.numeric(laptimes_data$lap[which.min(laptimes_data$time)])
25
26   cartella_pilota <- dirname(cartella_giro)
27   if(length(giro) > 0){
28     path_telemetry <- paste0(cartella_pilota, "/", giro, "_tel.json")
29   } else{
30     giro <- 1
31     path_telemetry <- paste0(cartella_pilota, "/", giro, "_tel.json") }
32
33   testo <- read_file(path_telemetry)
34   testo <- str_replace_all(testo, "NaN", "null")
35   json_data <- fromJSON(testo)[["tel"]]
36   json_data <- as_tibble(json_data)
37
38   parti_cartella <- str_split(cartella_giro, "/")[[1]]
39   n <- length(parti_cartella)
40
41   json_data$pilota <- parti_cartella[n-1]
42   json_data$GP <- parti_cartella[n-3]
43   json_data$lap_time <- rep(laptimes_data$time[giro], nrow(json_data))
44   lista_dati[[i]] <- json_data
45 }
46 #merge dei dati
47 dataset_completo <- bind_rows(lista_dati)
48 #pulizia
49 dataset_completo <- dataset_completo %>%
50 select(-dataKey)
51 setwd("C:/Users/feder/Documents/datasets/Computazionale/F1/data")
52 saveRDS(dataset_completo, file = "dataset_completo_best_tel.rds")
53
```

## B Codice R Commentato Esplorazione iniziale

```
1  # A seguito del caricamento delle librerie e del dataset si inizia l'esplorazione
   con le iniziali statistiche descrittive
2  summary(tel)
3  # Sistemazione delle variabili
4  tel$throttle <- ifelse(tel$throttle > 100, 100, tel$throttle)
5
6  #Trattamento degli NA e None
7
8  tel[which(is.na(tel$rel_distance)),]
9  tel <- tel %>% filter(pilota != "RUS" & GP != "Miami Grand Prix")
10
11  tel %>%
12  group_by(GP,pilota) %>%
13  filter(lap_time == "None") %>% summarize(n(),lap_time=max(lap_time))
14
15  tel <- tel %>% filter(lap_time != "None")
16
```

## C Codice R Commentato Trattamento variabili

```
1  tel.guida <- tel %>%
2  group_by(GP,pilota) %>%
3  mutate(
4  dec_x = if_else(acc_x < 0, abs(acc_x), 0),
5  acc_x = if_else(acc_x > 0, acc_x, 0),
6  acc_y = abs(acc_y)) %>%
7  ungroup()
8
```

## D Codice R Commentato Creazione variabili

```
1  #Creazione delle variabili lag
2  tel.guida <- tel.guida %>%
3  select(GP,pilota,throttle,acc_x,acc_y,dec_x,rel_distance) %>%
4  arrange(GP,pilota,rel_distance) %>%
5  group_by(GP,pilota) %>%
6  mutate(
7  across(
8  c(throttle, acc_x, acc_y,dec_x),
9  list(
10  lag1 = ~round(
11  ifelse(
12  lag(.x, 1) == 0 & .x == 0,
13  0,
14  ifelse(
15  lag(.x, 1) == 0,
16  (.x-0.01)/0.01,
17  (.x - lag(.x, 1)) / lag(.x, 1))
18  ),
19  4),
20  ...
21  ),
22  .names = "{.col}_{.fn}"
23  )
24  ) %>%
25  ungroup() %>%
26  select(-rel_distance)
27
```

## E Codice R Commentato Statistiche Riassuntive

```
1 tel.guida_summary <- tel.guida %>%
2 group_by(GP, pilota) %>%
3 {
4   lag_cols <- names(.) %>% .[str_detect(., "lag")]
5
6   summarise(.,
7     across(
8       c(throttle, acc_x, acc_y, dec_x),
9       list(
10        mean = ~round(mean(.x, na.rm = TRUE), 4),
11        sd = ~round(sd(.x, na.rm = TRUE), 4)
12      ),
13      .names = "{.col}_{.fn}"
14    ),
15
16    across(
17      all_of(lag_cols),
18      ~round(mean(.x[.x > 0], na.rm = TRUE), 4),
19      .names = "{.col}_mean_pos"
20    ),
21
22    across(
23      all_of(lag_cols),
24      ~round(mean(.x[.x <= 0], na.rm = TRUE), 4),
25      .names = "{.col}_mean_neg"
26    ),
27
28    across(
29      all_of(lag_cols),
30      ~round(sd(.x[.x > 0], na.rm = TRUE), 4),
31      .names = "{.col}_sd_pos"
32    ),
33
34    across(
35      all_of(lag_cols),
36      ~round(sd(.x[.x <= 0], na.rm = TRUE), 4),
37      .names = "{.col}_sd_neg"
38    ),
39
40    .groups = "drop"
41  )
42 }
43
44 tel2 <- tel %>%
45 group_by(GP, pilota) %>%
46 summarize(laptime=max(lap_time), strategy=(max(strategy)), gomme=(max(gomme))
47 , .groups = "drop")
48
49 tel.guida_summary$laptime <- as.numeric(tel2$laptime)
50 tel.guida_summary$strategy <- tel2$strategy
51 tel.guida_summary$gomme <- tel2$gomme
```



## F Codice R Commentato Analisi delle correlazioni

```
1  corr <- round(cor(tel.guida_summary %>% select(where(is.numeric))),4)
2
3  #Ricerca di variabili dipendenti
4
5  variabili_dipendenti <- findCorrelation(corr, cutoff = 0.9, names = TRUE, exact =
6    T,verbose = T)
7
8  print(variabili_dipendenti)
9
10 tel.pca <- tel.guida_summary %>% select(-all_of(variabili_dipendenti))
11
```

## G Codice R Commentato Analisi delle componenti principali

```
1  PCA <- princomp(tel.pca %>%
2    select(where(is.numeric)))
3
4  fviz_screplot(PCA,choice= "variance")
5
6  summary(PCA)
7  print(PCA$loadings,cutoff = 0)
8
9  for (i in 1:6){
10    p <- fviz_contrib(PCA, choice = "var", axes = i, top = 15)
11    print(p)
12  }
13
```