

# 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 che relative, vengono escluse dall'analisi.

In questa fase preliminare, lo stile di guida viene descritto attraverso variabili dinamiche, quali l'utilizzo dell'acceleratore, del freno, la velocità e le seguenti 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 e velocità per i piloti: Charles Leclerc, Lando Norris, Max Verstappen e Franco Colapinto.

---

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

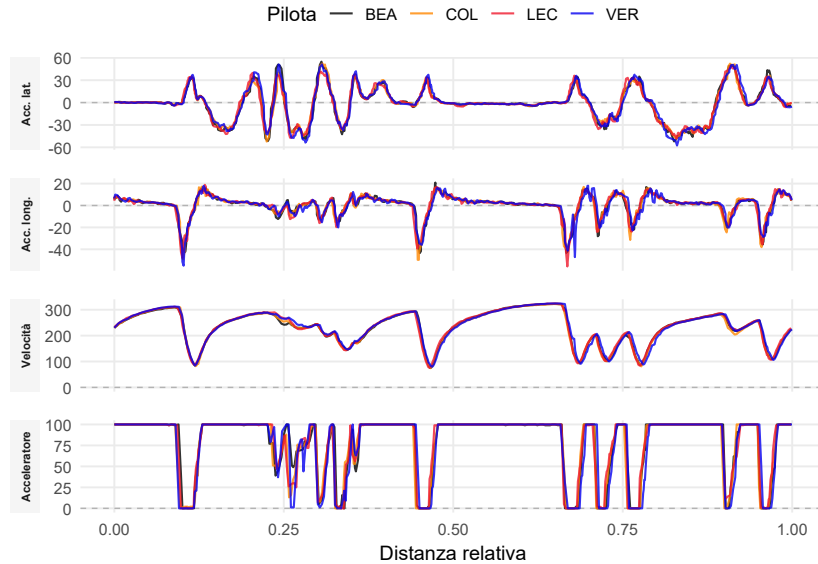


Figura 1: Accelerazioni VER, LEC, NOR, COL.

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.

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 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.

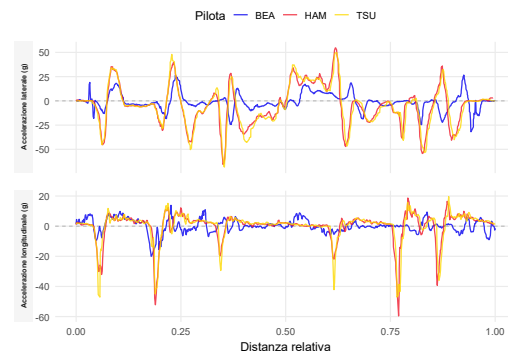


Figura 2: Confronto Australia Bearman

### 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;

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

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

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

Si creano per le variabili **velocità**, **acceleratore**, **decelerazione** e **accelerazioni**, 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, decelerazione e per gli input di frenata, accelerazione e velocità vengono calcolate media e deviazione standard.

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.

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 28 variabili, che forniscono informazione sullo stile di guida del pilota nella specifica gara.

Infine, per rimuovere l'influenza del tracciato, si riscalanano 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 26 variabili.

Seppur la dimensionalità sia diminuita, 26 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 4 componenti spiegano più del 60% della varianza, come si nota nella figura 3.

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

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

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

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

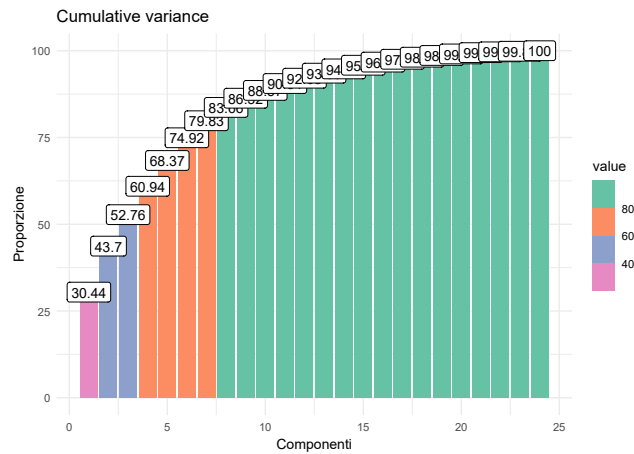


Figura 3: Varianza Spiegata Cumulata

### 2.3.1 Interpretazione delle componenti

Attraverso l'analisi dei pesi associati alle componenti (Figura 4) le si interpretano:

1. La prima componente ha pesi maggiori per le variabili che descrivono la variabilità della dinamica longitudinale, sono coinvolti i coefficienti di decelerazione ed accelerazione come anche le variazioni degli input come acceleratore, freno e velocità. A valori alti viene quindi associata una guida più instabile, per valori bassi invece è associata una guida più progressiva e stabile.
2. Per la seconda componente hanno pesi maggiori le variabili per lo più che descrivono la dinamica laterale insieme all'accelerazione longitudinale. A valori elevati è associata una maggiore variabilità, indicativa di intensità laterale più irregolare e impegnativa, spesso riguarda curve più accentuate. Valori bassi sono invece associati a una dinamica laterale più regolare e controllata spesso compatibile con la percorrenza di curve con andamento costante.
3. Nella terza componente prevalgono le dinamiche sia di accelerazione longitudinale che laterale, per quanto riguarda le laterali, i coefficienti che riguardano i lag negativi hanno maggiore peso, la componente descrive quindi la dinamica transitoria, per esempio un passaggio da curva a rettilineo. A valori alti viene associata una guida più instabile mentre per bassi valori viene associata una guida più progressiva e stabile.
4. Per la quarta componente i contributi più significativi sia della dinamica longitudinale, in particolare delle variabili di accelerazione e decelerazione con lag positivi, sia della dinamica laterale e della velocità. La componente descrive quindi fasi di guida dinamica in cui sono presenti variazioni coordinate di accelerazione, decelerazione e velocità, per esempio ingressi o uscite da curve. A valori elevati corrisponde una maggiore variabilità complessiva, associabile a uno stile di guida più dinamico e meno uniforme, mentre valori bassi indicano una guida più regolare e costante.

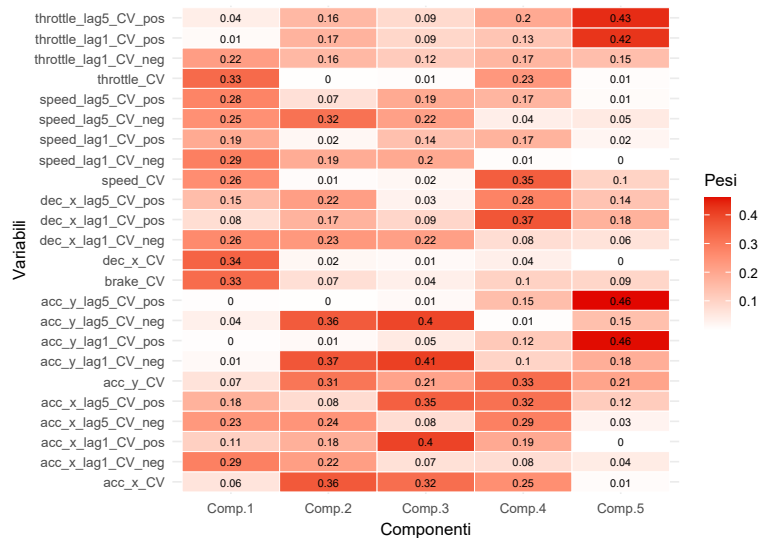


Figura 4: Pesi componenti

### 3 Model Based Clustering

L'analisi di Clustering Model-Based, condotta sulle prime quattro componenti principali, ha permesso di identificare tredici classi tramite un modello con vincoli su matrice di covarianza pari a; Volume variabile, forma sferica tra i gruppi stimando un totale di 77 parametri e  $BIC$  pari a  $-7108.435$ . Il raggruppamento risulta più distinguibile in due dimensioni confrontando graficamente prima componente con la seconda (figura 5a) anche se è presente del "rumore" grafico. Il che è dovuto per la maggior parte alle altre componenti come si nota nella figura 5b dove si distingue meglio il cluster 4 nel confronto grafico tra quarta, prima e seconda componente.

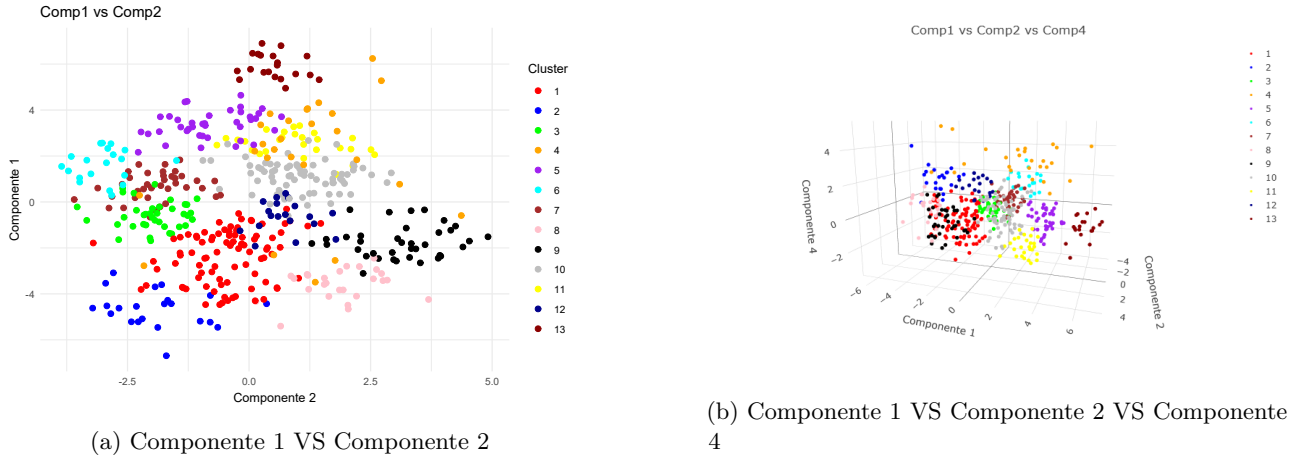


Figura 5: Confronto clustering

Dato che il modello ha un  $ICL$  pari a  $-7198.058$  confrontandolo con il  $BIC$  la distanza di Kullback-Leibler ( $KL$ ) risulta pari a 89.622 questo conferma che la struttura delle classi è ben distinta, il fatto viene confermato anche da un'incertezza attesa pari a 0.0796 si esclude quindi una possibile situazione di sovrapposizione dei gruppi.

Tabella 1: Media delle Componenti Principali per Classe

Classe	COMP 1	COMP 2	COMP 3	COMP 4	GP
1	-2.3544521	-0.4336353	-0.3267068	-1.04304992	Australian, Abu Dhabi, Qatar, Saudi, Spanish
2	-4.6485671	-1.9013751	0.3596806	1.50701769	Italian
3	-0.5809494	-1.9670107	0.3270949	-0.50006776	Austrian, Emilia Romagna
4	1.9782712	1.0854711	-1.4384612	3.35259093	Las Vegas
5	3.3611817	-0.7988630	0.3854497	-0.58053173	Mexico, Singapore
6	1.6093561	-2.9926369	3.1763503	1.47207130	Azerbaijan
7	0.6601369	-1.8975598	-1.5433976	0.33120784	Canadian, Bahrain
8	-3.5929136	1.8655423	1.9612523	0.21243282	Belgian
9	-1.7581531	3.1354654	0.7971459	-0.15900006	British, Japanese
10	1.0999736	1.0575898	-1.2098464	0.04385535	United States, São Paulo, Chinese, Dutch
11	2.5236924	1.0020324	-0.4849935	-1.88106300	Hungarian, Dutch
12	-0.6392196	0.7454101	0.5008869	1.58845641	Miami
13	5.8976535	0.5075088	1.8048374	-0.47820303	Monaco

- Classe 13: Questa classe identifica univocamente il Gran premio di Monaco. Il valore massimo della prima componente infatti si può associare ad un uso del gas continuo;
- Classi 5 e 11: (Singapore, Messico, Ungheria) Queste classi raggruppano circuiti tortuosi ad alto carico. Condividono con Monaco la necessità di gestire l'acceleratore per gestire per esempio un uscita da una curva netta.
- La Classe 6: (Baku) mostra un'alta terza componente: le curve a  $90^\circ$  costringono correzioni veloci della accelerazione laterale.
- La Classe 4: (Las Vegas) e 12 (Miami) mostrano un'alta quarta componente: l'asfalto scivoloso a soprattutto a Las Vegas per la pioggia creano incertezza nel mantenere la velocità costante, costringendo a parzializzare il gas dove normalmente si andrebbe costanti.

- Classi 8 e 9: (Spa, Silverstone, Suzuka) Queste classi rappresentano i circuiti con prima componente fortemente negativa non c'è "nervosismo" sul gas dovuto anche alle curve più fluide dei tracciati. La seconda componente positiva distingue questi circuiti: indica che la variazione è dominata dalle forze longitudinali ad alta velocità.
- Classe 10: (Cina, USA, Brasile) Raggruppa circuiti molto tecnici. La prima componente positiva qui non indica instabilità, ma la parzializzazione del gas dovuta a curve differenti sul tracciato. La terza componente negativa conferma che, nonostante la difficoltà tecnica, l'esecuzione è stabile, senza le correzioni tipiche dei cittadini.
- Monza (Classe 2) e Bahrain/Canada (Classe 7): la prima componente molto negativa. Sono piste di staccata profonda con un'uscita di trazione pulita, seconda componente negativa.
- Imola (Classe 3) e Spagna (Classe 1): seconda componente negativa e prima bassa. Qui prevale la percorrenza e la gestione del carico laterale.

## 4 Classificazione

## 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 qualifiche
8 file_giri <- list.files(path = cartella,
9 pattern = "laptimes\\.json$",
10 full.names = TRUE,
11 recursive = TRUE)
12
13 files_qualifiche_lap <- file_giri[grepl("Qualifying", file_giri) & !grepl("Sprint",file_
14 giri)]
15 lista_dati <- list()
16
17 #Ciclo di estrazione
18 for(i in 1:length(files_qualifiche_lap)) {
19
20   cartella_giro <- files_qualifiche_lap[i]
21
22   laptimes_data <- fromJSON(cartella_giro)
23
24   giro <- as.numeric(laptimes_data$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
2 # iniziali statistiche descrittive
3 summary(tel)
4 # Sistemazione delle variabili
5 tel$throttle <- ifelse(tel$throttle > 100, 100, tel$throttle)
6
7 #Trattamento degli NA e None
8
9 tel[which(is.na(tel$rel_distance)),]
10 tel <- tel %>% filter(pilota != "RUS" & GP != "Miami Grand Prix")
11
12 tel %>%
13 group_by(GP,pilota) %>%
14 filter(lap_time == "None") %>% summarize(n(),lap_time=max(lap_time))
15
16 tel <- tel %>% filter(lap_time != "None")
```

## 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)), gomma=(max(gomma)), .groups = "
47 drop")
48
49 tel.guida_summary$laptime <- as.numeric(tel2$laptime)
50 tel.guida_summary$strategy <- tel2$strategy
51 tel.guida_summary$gomma <- tel2$gomma
```

## 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 = T, verbose
6   = 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
```