GRUPPO: CORES

STATISTICA COMPUTAZIONALE – REPORT FINALE

Maccianti Federico (909656) Rapacioli Nicola (915439) Riva Pietro (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[1]) 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	–	$\{1, \ldots, 8\}$	`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.

[1] Codice di estrazione dati nell'appendice A.

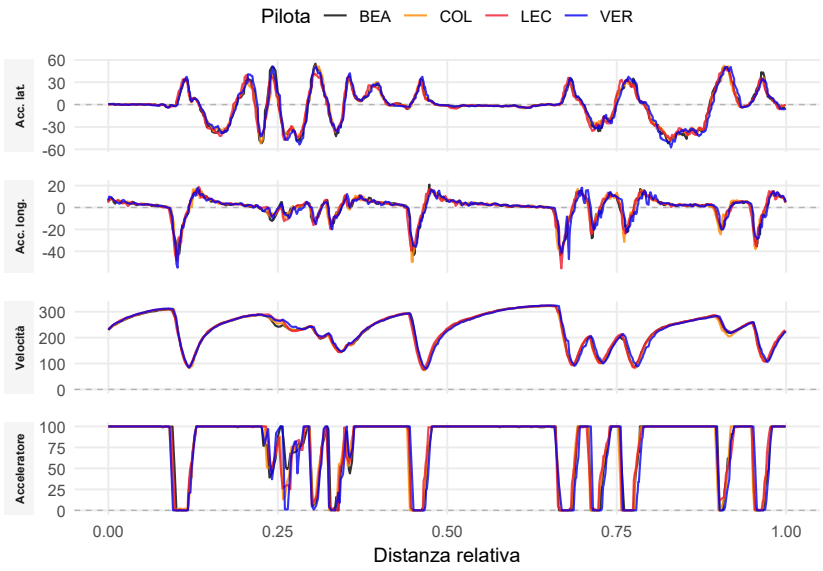


Figura 1: Accelerazioni VER, LEC, NOR, COL.

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

Dalle prime analisi descrittive sul dataset[2], 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[2], 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.

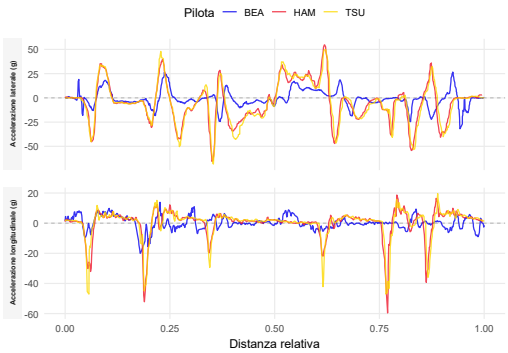


Figura 2: Confronto Australia Bearman

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 Bortoleto 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)[3].

Inoltre, l'`accelerazione longitudinale` viene suddivisa[3] in:

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

[2]Estratto codice R consultabile nell'appendice B.
[3]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[4], 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[5] 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 `_lag`k 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 cosi 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 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

Al fine di limitare i fenomeni di multicollinearità ed eliminare il rumore, sono state eliminate le variabili fortemente correlate tra loro ($cor(x, y) > 0.9$)[6]. Questo processo di scrematura permette di ridurre la dimensionalità a 26 variabili.

Nonostante tale riduzione, una dimensionalità pari a 26 risulta ancora essere sia ai fini di interpretabilità che per costo computazionale per un Model Base Clustering.

Si effettua dunque un analisi delle componenti principali (PCA) con l'obiettivo di ottenere un'ulteriore riduzione dimensionale. I risultati ottenuti tramite PCA[7] mostrano come le prime 4 componenti spieghino più del 60% della varianza, come illustrato in figura 3.

[4] Estratto codice R consultabile nell'appendice D
[5] Estratto codice R consultabile nell'appendice E.
[6] Estratto del codice R consultabile nell'appendice F
[7] Estratto del codice R consultabile nell'appendice G

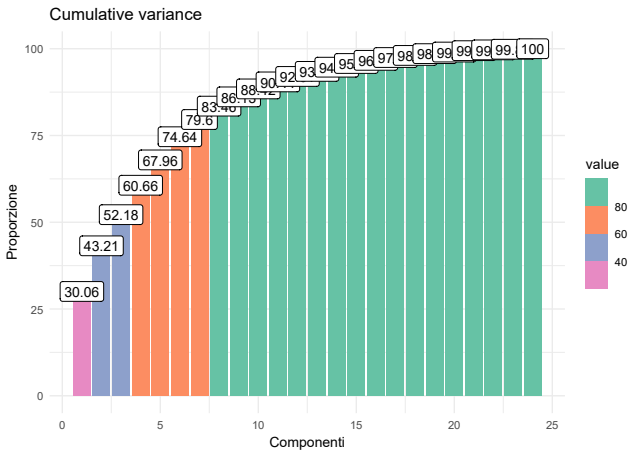


Figura 3: Varianza Spiegata Cumulata

2.3.1 Interpretazione delle componenti

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

1. **Prima componente** : Separa stili stili di guida che hanno un'alta variabilità (molte correzioni) nell'erogazione della potenza da quelli che hanno un'alta variabilità nella fase di decelerazione. A valori elevati corrisponde a uno stile di guida caratterizzato da una trazione più "sporca" o reattiva, con continue correzioni sul pedale dell'acceleratore che si contrappone uno stile caratterizzato da una frenata molto modulata e dinamica, mantenendo però un'uscita di curva molto composta e lineare nell'erogazione del gas.

2. **Seconda componente** : distungue stili di guida che lavorano di più in termini di velocità pura (avanti/-dietro) contro chi ha maggiore di percorrenza (destra/sinistra). A valorio elevati corrisponde uno stile di guida a "V", dove il pilota punta tutto sulla frenata profonda e sulla ripartenza rapida, variando molto l'accelerazione longitudinale. Invece, per valori bassi contrappone uno stile di guida basato sulla velocità di percorrenza, che predilige la gestione del carico laterale.

3. **Terza componente** : questa componente entra nel dettaglio delle fasi della curva. Discrimina stili di guida con bassa componente laterale in fase di percorrenza e alta componente di accellerazioni longitudinali e viceversa. In particolare, a valori alti corrisponde uno stile di guida caratterizzato forti staccate, forti accellerazioni in uscita curva e bassa velocità di percorrenza. Mentre a valori bassi corrisponde uno stile caratterizzato alte velocità di percorrenza, e poca variazione longitudinale.

4. **Quarta componente** : questa componente caratterizza la configurazione della pista. A valori elevati corrisponde una maggiore variabilità complessiva della velocità, mentre valori bassi indicano un'andamento più regolare e uniforme.

N.B. Il circuito vincola il pilota a modificare il proprio stile di guida alla configurazione del tracciato, alternando fluidità e aggressività in base alle specifiche richieste aerodinamiche o meccaniche. Tale adattamento serve quindi esclusivamente a ottimizzare lo sfruttamento della massima aderenza disponibile in ogni circuito.

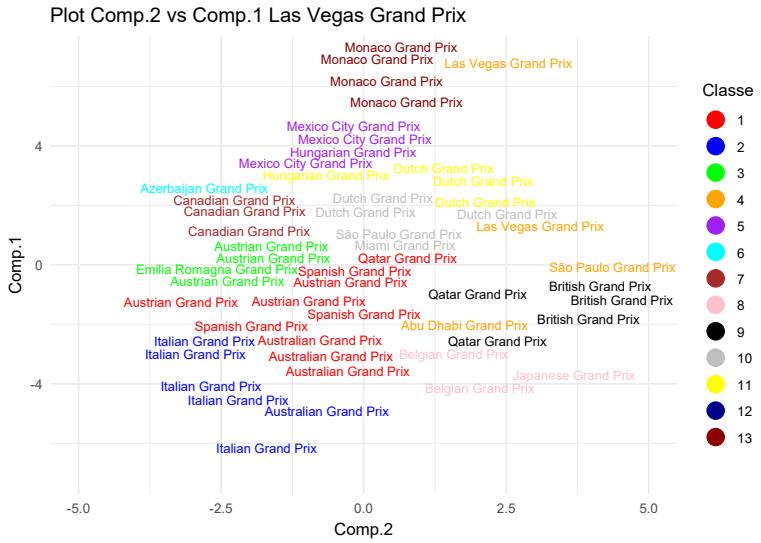
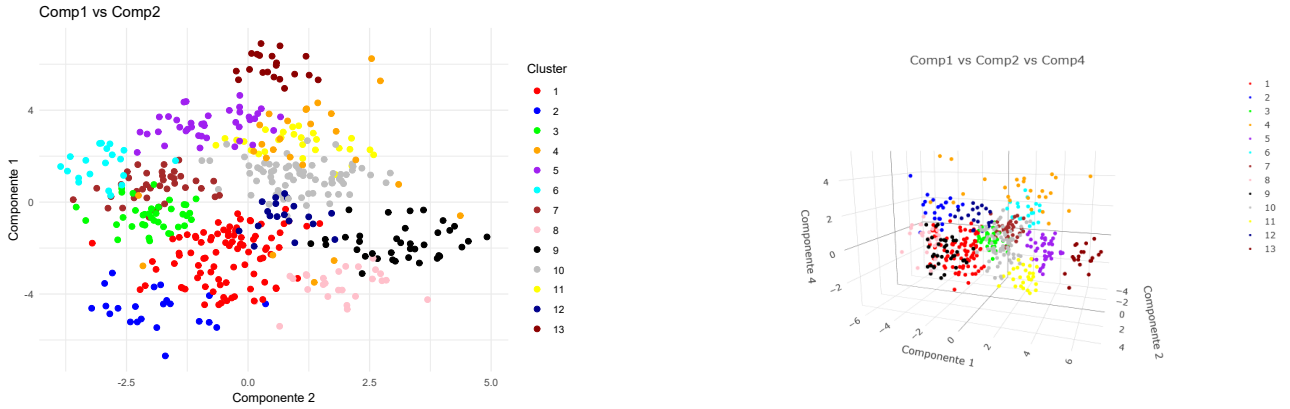


Figura 4: Pesi componenti

3 Model Based Clustering

L'analisi di Clustering Model-Based, condotta sulle prime quattro componenti, ha permesso di identificare un modello *VII* (*Volume variabile*, *forma sferica* e *orientamento identico*) a tredici classi, con un totale di 77 parametri stimati e BIC pari a -7108.435.

Il raggruppamento dei dati risulta già chiaramente distinguibile nella rappresentazione bidimensionale ottenuta dal confronto tra la prima e la seconda componente principale (Figura 5a), sebbene sia ancora presente una certa sovrapposizione visiva, interpretabile come "rumore" grafico. Tale effetto è attribuibile principalmente all'informazione contenuta nelle componenti principali rimanenti. Ciò è ulteriormente evidenziato dalla Figura 5b, in cui il confronto congiunto tra la quarta, la prima e la seconda componente consente di distinguere in modo più netto i cluster, migliorandone la separabilità rispetto alla rappresentazione bidimensionale.



(a) Componente 1 VS Componente 2

(b) Componente 1 VS Componente 2 VS Componente 4

Figura 5: Confronto clustering

Il valore della distanza di Kullback-Leibler (KL) del modello risulta essere pari a 89.622. Ciò dà sostegno a quanto già osservato graficamente, ulteriormente supportato dall'incertezza pari a 0.0796 che esclude una possibile situazione di sovrapposizione dei gruppi.

- Classe 13: Questa classe identifica uno stile il Gran premio di Monaco. Il valore molto elevato della prima componente infatti si può associare ad un uso del gas ritmico.
- 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° costringono correzioni veloci della accelerazione laterale.

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

Tabella 1: Media delle Componenti Principali per Classe

- La Classe 4 e 12: (Las Vegas e 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 dovuta all'assenza di frenate decise. 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 completi dal punto di vista del tracciato : alternano tratti veloci a tratti tecnici. Non risulta esserci una componente con pesi notevoli.
- Classe 2 : (Monza): La prima componente molto negativa. È una pista caratterizzata da staccate profonde con un'uscita di trazione pulita, inoltre la quarta componente abbastanza alta, dovuta alle elevate variazioni di velocità.
- Classe 7 : (Bahrain e Canada): Sono circuiti caratterizzate da poche curve di alta velocità di percorrenza, evidenziato dal valore negativo della seconda componente.
- Classe 1 : (Australian, Abu Dhabi, Qatar, Saudi, Spanish) È il gruppo più numeroso, caratteizzato da una prima componente negativa, quindi poche variazioni longitudinali e in generale fluidità nella percorrenza.
- Classe 3 : (Austrian, Emilia Romagna) Molto simile alla classe 1, ma con curve più accentuate e maggiori variazioni longitudinali.

4 Classificazione

A Codice R Commentato Estrazione dei dati

```r
library(jsonlite)
library(tidyverse)
rm(list=ls())
#Cartella origine
cartella <- "C:/Users/feder/Documents/datasets/Computazionale/2025"

#filtro sui file con i giri e poi qualifiche
file_giri <- list.files(path = cartella,
pattern = "laptimes\\.json$",
full.names = TRUE,
recursive = TRUE)

files_qualifiche_lap <- file_giri[grepl("Qualifying", file_giri) & !grepl("Sprint",file_
giri)]
lista_dati <- list()

#Ciclo di estrazione

for(i in 1:length(files_qualifiche_lap)) {

  cartella_giro <- files_qualifiche_lap[i]

  laptimes_data <- fromJSON(cartella_giro)

  giro <- as.numeric(laptimes_data$lap[which.min(laptimes_data$time)])

  cartella_pilota <- dirname(cartella_giro)
  if(length(giro) > 0){
    path_telemetry <- paste0(cartella_pilota, "/", giro, "_tel.json")
  } else{
    giro <- 1
    path_telemetry <- paste0(cartella_pilota, "/", giro, "_tel.json") }

  testo <- read_file(path_telemetry)
  testo <- str_replace_all(testo, "NaN", "null")
  json_data <- fromJSON(testo)[["tel"]]
  json_data <- as_tibble(json_data)

  parti_cartella <- str_split(cartella_giro, "/")[[1]]
  n <- length(parti_cartella)

  json_data$pilota <- parti_cartella[n-1]
  json_data$GP <- parti_cartella[n-3]
  json_data$lap_time <- rep(laptimes_data$time[giro],nrow(json_data))
  lista_dati[[i]] <- json_data
}
#merge dei dati
dataset_completo <- bind_rows(lista_dati)
#pulizia
dataset_completo <- dataset_completo %>%
select(-dataKey)
setwd("C:/Users/feder/Documents/datasets/Computazionale/F1/data")
saveRDS(dataset_completo, file = "dataset_completo_best_tel.rds")

```

B Codice R Commentato Esplorazione iniziale

```r
# A seguito del caricamento delle librerie e del dataset si inizia l'esplorazione con le iniziali statistiche descrittive
summary(tel)
# Sistemazione delle variabili
tel$throttle <- ifelse(tel$throttle > 100, 100, tel$throttle)

#Trattamento degli NA e None

tel[which(is.na(tel$rel_distance)),]
tel <- tel %>% filter(pilota != "RUS" & GP != "Miami Grand Prix")

tel %>%
group_by(GP,pilota) %>%
filter(lap_time == "None") %>% summarize(n(),lap_time=max(lap_time))

tel <- tel %>% filter(lap_time !="None")

```

C Codice R Commentato Trattemento variabili

```r
tel.guida <- tel %>%
group_by(GP,pilota) %>%
mutate(
dec_x = if_else(acc_x < 0, abs(acc_x), 0),
acc_x = if_else(acc_x > 0, acc_x, 0),
acc_y = abs(acc_y)) %>%
ungroup()

```

D Codice R Commentato Creazione variabili

```r
#Creazione delle variabili lag
tel.guida <- tel.guida %>%
select(GP,pilota,throttle,acc_x,acc_y,dec_x,rel_distance) %>%
arrange(GP,pilota,rel_distance) %>%
group_by(GP,pilota) %>%
mutate(
across(
c(throttle, acc_x, acc_y,dec_x),
list(
lag1 = ~round(
ifelse(
lag(.x, 1) == 0 & .x == 0,
0,
ifelse(
lag(.x, 1) == 0,
(.x-0.01)/0.01,
(.x - lag(.x, 1)) / lag(.x, 1))
),
4),
...
),
.names = "{.col}_{.fn}"
)
)  %>%
ungroup() %>%
select(-rel_distance)

```

E Codice R Commentato Statistiche Riassuntive

```r
tel.guida_summary <- tel.guida %>%
group_by(GP, pilota) %>%
{
  lag_cols <- names(.) %>% .[str_detect(., "lag")]

  summarise(.,
  across(
  c(throttle, acc_x, acc_y,dec_x),
  list(
  mean = ~round(mean(.x, na.rm = TRUE), 4),
  sd = ~round(sd(.x, na.rm = TRUE), 4)
  ),
  .names = "{.col}_{.fn}"
  ),

  across(
  all_of(lag_cols),
  ~round(mean(.x[.x > 0], na.rm = TRUE), 4),
  .names = "{.col}_mean_pos"
  ),

  across(
  all_of(lag_cols),
  ~round(mean(.x[.x <= 0], na.rm = TRUE), 4),
  .names = "{.col}_mean_neg"
  ),

  across(
  all_of(lag_cols),
  ~round(sd(.x[.x > 0], na.rm = TRUE), 4),
  .names = "{.col}_sd_pos"
  ),

  across(
  all_of(lag_cols),
  ~round(sd(.x[.x <= 0], na.rm = TRUE), 4),
  .names = "{.col}_sd_neg"
  ),

  .groups = "drop"
  )
}

tel2 <- tel %>%
group_by(GP,pilota) %>%
summarize(laptime=max(lap_time),strategy=(max(strategy)),gomma=(max(gomma)),.groups = "
drop")

tel.guida_summary$laptime <- as.numeric(tel2$laptime)
tel.guida_summary$strategy <- tel2$strategy
tel.guida_summary$gomma <- tel2$gomma

```

F Codice R Commentato Analisi delle correlazioni

```r
corr <- round(cor(tel.guida_summary %>% select(where(is.numeric))),4)

#Ricerca di variabili dipendenti

variabili_dipendenti <- findCorrelation(corr, cutoff = 0.9, names = TRUE, exact = T,verbose
  = T)

print(variabili_dipendenti)

tel.pca <- tel.guida_summary %>% select(-all_of(variabili_dipendenti))


```

G Codice R Commentato Analisi delle componenti principali

```r
PCA <- princomp(tel.pca %>%
select(where(is.numeric)))

fviz_screeplot(PCA,choice= "variance")

summary(PCA)
print(PCA$loadings,cutoff = 0)

for (i in 1:6){
  p <- fviz_contrib(PCA, choice = "var", axes = i, top = 15)
  print(p)
}

```