# Analisi del Dataset *Human Resources*

Corso di Programmazione e Analisi di Dati

D'Antoni Giorgia (620020), Cascione Alessio (582765)
Professore: Alessio Malizia
a.a. 2021-2022

# Indice

# 1. Data Understanding

Il dataset analizzato prende il nome di HR Dataset v14, ed è un dataset per le risorse umane composto da 311 righe e 36 colonne, anche dette *features* del nostro dataset, contenenti informazioni riguardo una società fittizia.
La nostra attenzione sarà focalizzata, in particolare, sul rapporto tra tre variabili principali da noi scelte e le altre variabili presenti nel dataset, le tre variabili principali prese in considerazione sono: il salario, il livello di performance degli impiegati e le assenze.
Attraverso l'utilizzo di un grafico a torta per osservare la percentuale con cui uomini e donne sono presenti nel dataset, possiamo notare che le due variabili sono presenti in maniera non bilanciata nel dataset: le donne sono infatti presenti in maggiore quantità rispetto agli uomini con una percentuale del 56.6 %.

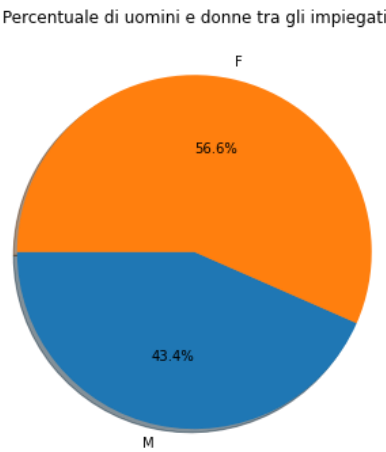


*Figura 1 – Percentuale di uomini e donne nel dataset*

## 1.1 Valori Mancanti

La presenza di valori mancanti all'interno di un dataset può essere riconducibile a diverse cause, come un errore nella compilazione, un rifiuto nel rispondere, un sensore rotto o altro. In questo caso le variabili presentano rispettivamente 207 valori mancanti per *DateofTermination* e soltanto 8 per *ManagerID*. I valori mancanti di *DateofTermination* non sono legati ad errori di inserimento o mancanza di informazione, ma la semantica della variabile prevede che questa non assuma valore per quegli impiegati ancora attivi: da ciò deriva l'alto numero di valori mancanti per la feature in questione e la decisione di non trattare i valori mancanti.

## 1.2 Analisi statistica

| | Salario | Progetti | Assenze | Engagement | Ritardi |
|---|---|---|---|---|---|
| *Skewness* | 3.306181 | 1.539271 | 0.029283 | -1.116979 | 3.143468 |
| *Kurtosis* | 15.452149 | 0.641415 | -1.301962 | 1.164560 | 8.830523 |
| Media | 69020.7 | 1.21 | 10.23 | 4.11 | 0.41 |
| Deviazione standard | 25156.63 | 2.35 | 5.85 | 0.79 | 1.29 |

*Tabella 1 – Principali indici statistici per variabili quantitative*

Per valutare l'affidabilità o meno della media della distribuzione, consideriamo la seguente euristica: se la deviazione standard della distribuzione è inferiore al 30% del valore della media, giudicheremo affidabile il valore medio della distribuzione. In caso contrario verrà valutato come non rappresentativo. Dai calcoli si è concluso che solo per la variabile *EngagementSurvey* abbiamo un valore medio affidabile. Negli altri casi quindi si potrebbe optare per scegliere la mediana come valore centrale rappresentativo della distribuzione.
Utilizzando la misura statistica *Kurtosis* per sapere quanto è alto e acuto il picco della distribuzione di una variabile, possiamo osservare una distribuzione fortemente pronunciata verso l'alto per le variabili *Salary* (15.4521) e *DaysLateLast30* (8.8305), *Absences* presenta invece una tendenza opposta con un valore di 0.0292 indice di una campana maggiormente schiacciata. Per quanto riguarda la skewness della curva è possibile avanzare alcune osservazioni: per *Absences* abbiamo un ottimo valore di inclinazione, indicando che i dati si distribuiscono in modo abbastanza simmetrico, a differenza di *DaysLateLast30* (3.1434) e *Salary* (3.3061) che presentano valori particolarmente alti per i quali ci aspetteremmo un forte sbilanciamento a favore della porzione della distribuzione con valori minori.
Analogamente, *SpecialProjectsCount* presenta una moderata distorsione dei dati verso valori bassi, mentre *EngagementSurvey* una moderata distorsione verso valori più alti.
Per una migliore comprensione delle conclusioni tratte fino ad adesso, visualizziamo le curve di distribuzione di ciascuna colonna presa in considerazione:

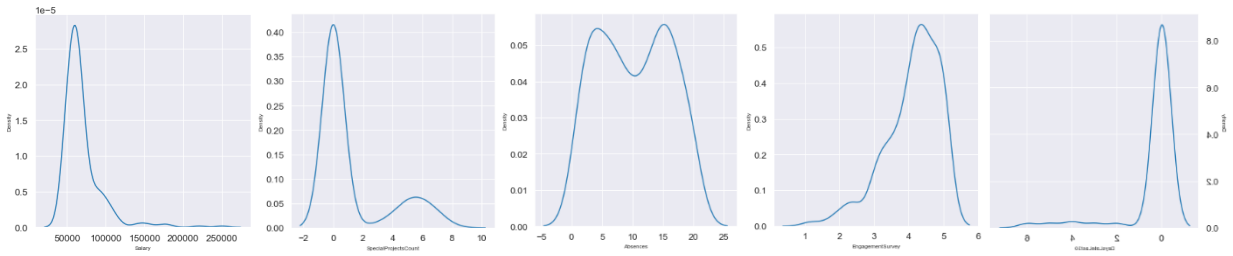

*Figura 2 – Curva di distribuzione: Salary, SpecialProjectsCount, Absences, EngagementSurvey, DaysLateLast30*

In nessun caso si può dunque parlare di distribuzione normale: anche se alcune variabili presentano un basso livello di Kurtosis, non vantano invece valori soddisfacenti per l'inclinazione della curva e viceversa. Per valutare dunque il livello di correlazione tra le variabili occorrerebbe prediligere test di correlazione non parametrici.

## 1.3 Individuazione e trattamento degli Outliers

Proseguiamo approfondendo il calcolo dei percentili per le variabili quantitative e sfruttiamo il range interquartile per individuare potenziali outliers delle distribuzioni: dividiamo gli outliers individuati in *upper-outliers* (che assumono, rispetto al terzo quartile, un valore maggiore di 1.5 volte il range interquartile) e *lower-outliers* (che assumono, rispetto al primo quartile, un valore minore di 1.5 volte il range interquartile). L'unica variabile che non presenta alcun tipo di outliers, considerando il range interquartile, è *Absences*, per il resto, notiamo come la maggior parte degli outliers si concentrano oltre il terzo quartile con 29 outliers in *Salary*, 70 in *SpecialProjectsCount*, 30 in *DaysLateLast30*. Soltanto *EngagementSurvey* presenta 9 outliers inferiori al primo quartile. Rimuovendo gli outliers otteniamo un avvicinamento maggiore della media (62841) al valore mediano (61620) ed una sosanziale riduzione della deviazione standard per quanto riguarda la variabile *Salary*.

## 1.4 Normalizzazione

Al fine di migliorare la distribuzione dei dati, evitando dunque che vi siano dei valori molto più alti o più bassi degli altri, possono essere applicati due diversi metodi di normalizzazione: *zscore e MinMaxScaler*. La normalizzazione attraverso *zscore* consiste nel sottrarre la media dai diversi valori presi in input e dividere il risultato ottenuto per la loro deviazione standard. I risultati ottenuti mostrano valori negativi e positivi: uno zscore negativo sta ad indicare un valore iniziale sotto la media, mentre un valore positivo sta ad indicare un valore iniziale sopra la media. Infatti, si può stabilire che, ad esempio, nella seconda riga della variabile *SpecialProjectsCount*, che riporta un valore di 2.0383, il valore originale è di 2 deviazioni standard sopra la media. Il metodo *MinMaxScaler*, diversamente da *zscore*, scala i valori in un range che va da 0 ad 1: dove lo zero indica il valore minimo di ogni colonna e l'uno il valore massimo. In questo modo tutti i valori rientrano nel range prestabilito, impedendo ai valori esageratamente alti di avere un peso maggiore rispetto agli altri dati.

## 2. Analisi del Salario

Ci focalizziamo ora sul modo in cui il salario è distribuito rispetto alle diverse unità statistiche del dataset: visualizziamo prima il comportamento della variabile facendo uso di un istogramma, con lo scopo di osservare in generale la distribuzione di questa. Ci concentreremo poi su uno studio mirato del salario rispetto ad altre categorie. Le linee tratteggiate enfatizzano due principali indici di tendenza centrale: la media (69020.70 dollari), evidenziata in rosso, e la mediana (62810 dollari), in verde, più vicina al picco della distribuzione continua della variabile.

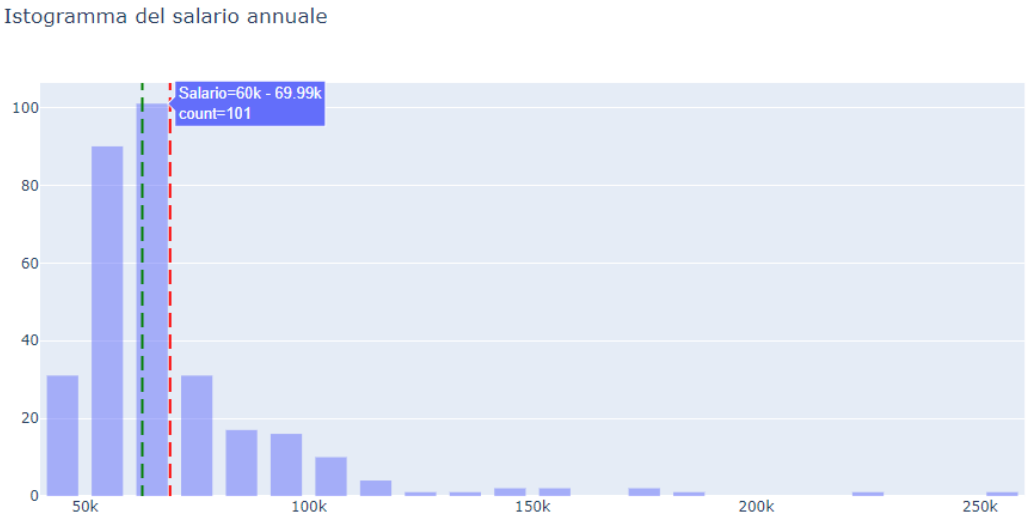


*Figura 3- Istogramma dei livelli di salario*

Focalizzando l' attenzione sul modo in cui il salario è distribuito in classi specifiche del dataset, è stato utilizzato un boxplot (figura 4) per visualizzare come valori diversi di salario si distribuiscano per le differenti etnie, consentendoci di individuare valori anomali: è interessante evidenziare come i dati che rientrano nell'intervallo tra primo e terzo quartile per l'etnia *White* tendano ad essere distribuiti più uniformemente, se consideriamo la mediana del boxplot di quella categoria, rispetto alle altre due etnie di maggioranza, *Asian* e *Black or African America*n. È altrettanto interessante notare come sia proprio *White* l'etnia che presenta il maggior numero di outliers nelle posizioni lavorative di *Database Administrator* e *Software Engineer*, nel gruppo rientra anche colei che ricopre il ruolo di *President & CEO*. Gli outliers del gruppo in questione sono anche uniformemente distribuiti dal punto di vista del genere, con 11 donne e 10 uomini in totale. Al contrario, invece, dei 6 outliers che caratterizzano il gruppo *Black or African American*, di cui solo uno è di sesso femminile mentre i restanti 5 sono uomini. In questo caso, la maggior parte degli impiegati del gruppo ricopre diversi ruoli identificati sotto la posizione di *IT Manager*.

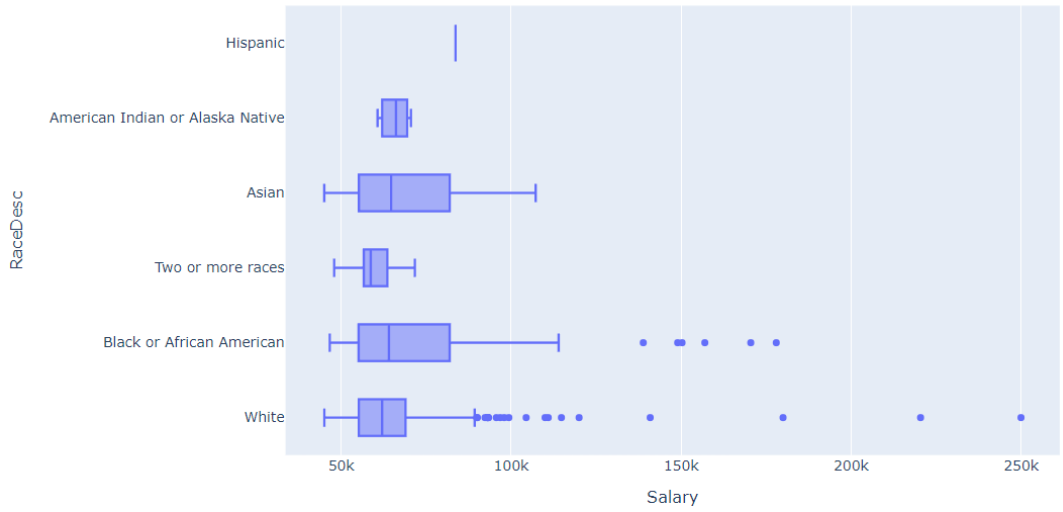


*Figura 4- Distribuzione del salario nelle diverse etnie*

Ci focalizziamo successivamente sui livelli di salario medio rispetto alle etnie con particolare attenzione sui due generi: per tutte le etnie, escludendo *White*, gli impiegati di sesso maschile hanno uno stipendio medio maggiore di quelli di sesso femminile. La differenza è piuttosto evidente per la categoria *Black or African American*, dove lo stipendio medio femminile equivale a 66963.83 dollari mentre quello maschile è pari a 85066.12. La categoria *Hispanic* presenta invece solo un individuo di sesso maschile ed è quindi l'etnia meno rappresentata nel dataset (figura 5).

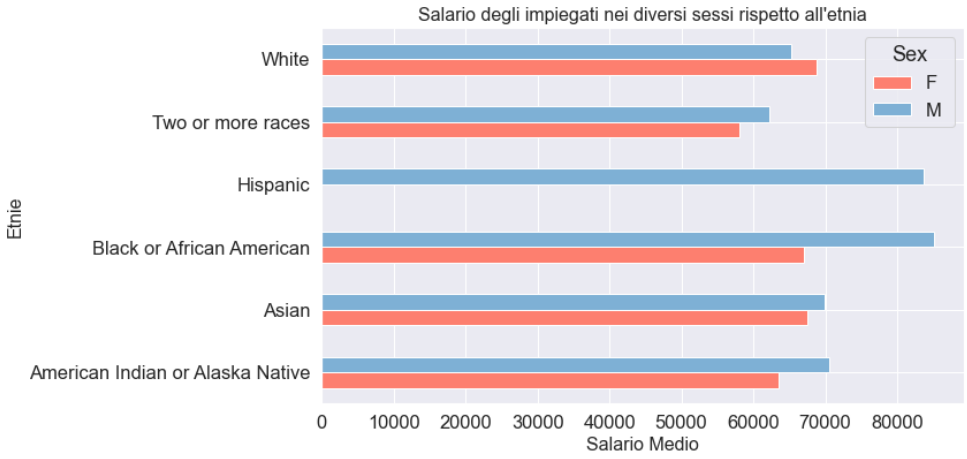


*Figura 5- Salario degli impiegati di diversa etnia con focus sul genere*

La figura 6, invece, mostra il livello di salario percepito dai diversi generi a seconda delle posizioni lavorative ricoperte. Risulta evidente che *President & CEO* e *Chief Information Officer* siano le posizioni che prevedono un salario maggiore e sono ricoperte da due impiegati di sesso femminile. Al contrario, le posizioni di tipo

informatico, come tutte quelle relative a *IT Manager*, tendono ad essere ricoperte principalmente da individui di sesso maschile.

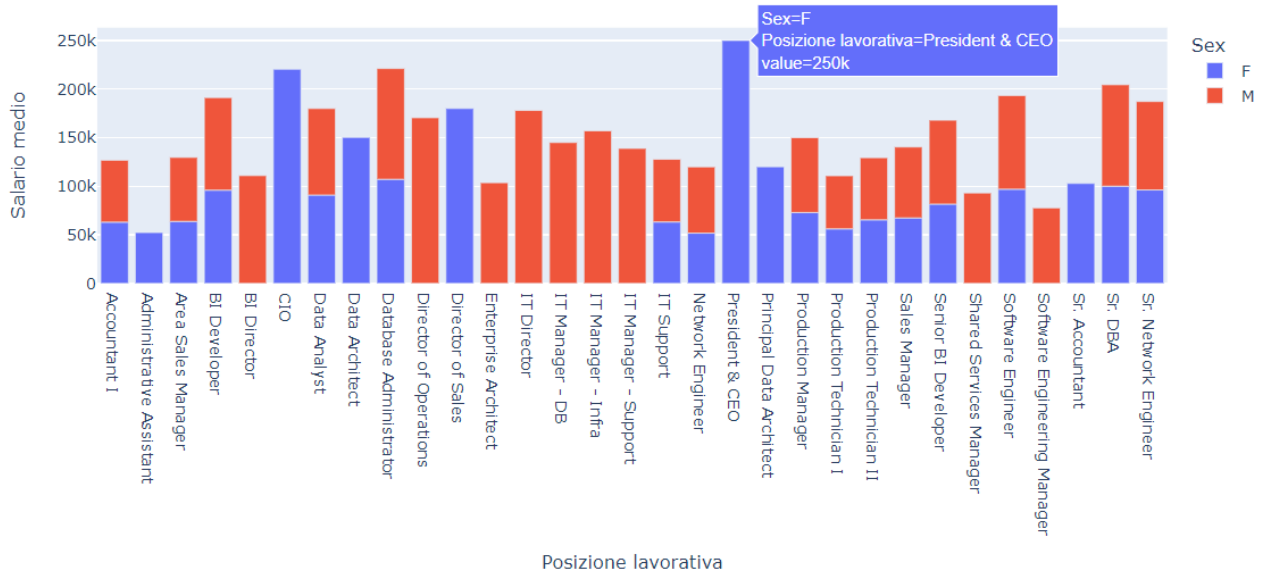


*Figura 6- Salario per posizione lavorativa con focus sul genere*

## 3. Analisi della Performance

Successivamente spostiamo il nostro focus sull'analisi della performance di ogni persona presente nel nostro dataset. Fonte di interesse è comprendere come le altre variabili si dividono nelle diverse performance attraverso l'utilizzo di grafici (barplot). Per avere una più chiara rappresentazione della variabile *PerformanceScore* al quale faremo riferimento la visualizziamo attraverso un piechart:

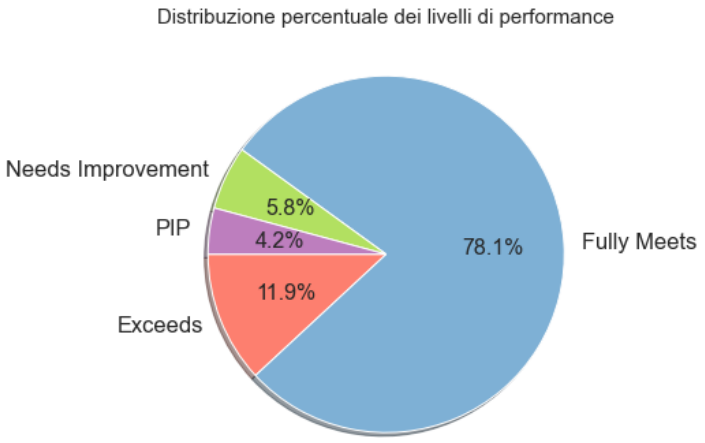


*Figura 7- Livelli di performance in percentuale nel Dataset*

Attraverso l'utilizzo del grafico è possibile osservare che la variabile è composta da 4 diversi livelli di performance: *PIP*, *Needs Improvement*, *Fully Meets*, *Exceeds*. I primi due livelli menzionati possono definirsi i più "bassi", ovvero con un grado di performance non ottimale e, come visibile dalle diverse percentuali presenti sul grafico, sono anche quelli in minoranza con una percentuale di 4,18% per *PIP* ed una percentuale di 5,79% per *Needs Improvement*. Il livello di performance che prevale sugli altri è evidentemente il *Fully Meets* presente in una percentuale del 78,18%, implicando che la maggior parte degli impiegati si impegna nel proprio lavoro con una performance sufficiente per gli standard imposti dall'azienda. Infine, possiamo vedere che soltanto l'11,90% delle persone presenti nel dataset eccede nel proprio lavoro.
Una volta visualizzata in percentuale la composizione della variabile *PerformanceScore*, si possono utilizzare dei barplot per osservare come altre variabili presenti nel dataset si dividono a seconda dei diversi livelli di performance precedentemente menzionati: un primo barplot è stato creato per visualizzare la composizione delle diverse tipologie di performance in base al sesso dell'impiegato, rendendo visibile una parità di genere quasi in ogni livello di performance, con una presenza lievemente maggiore di individui di sesso femminile nelle classi *Exceeds* e *Fully Meets*, e una presenza lievemente maggiore di individui di sesso maschile nella classe *PIP*. Il grafico ci potrebbe dunque portare a dedurre che in media le donne tendono ad impegnarsi maggiormente degli uomini, ma bisogna anche tenere da conto che le classi *F* e *M* della variabile *Sex* siano presenti in forma sbilanciata dal momento che il numero degli individui di sesso femminile supera quello degli individui di sesso maschile.

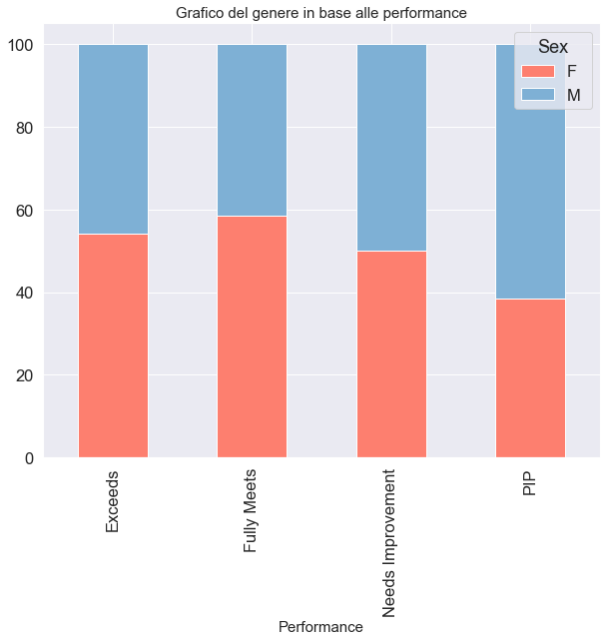


*Figura 8- Livello di performance in base al genere*

Visualizziamo poi come la variabile *MaritalDesc* si suddivide nei diversi livelli di performance. Risulta evidente che le classi *Single* e *Married* prevalgono in tutti i livelli di performance, con un'elevata presenza in *Fully Meets* e *Exceeds*, probabilmente derivante dal fatto che le persone sposate possiedono spesso famiglia e dunque non possono rischiare di perdere il lavoro o guadagnare poco. Allo stesso modo le persone single, non avendo ulteriori distrazioni, tendono a focalizzarsi spesso sul lavoro.

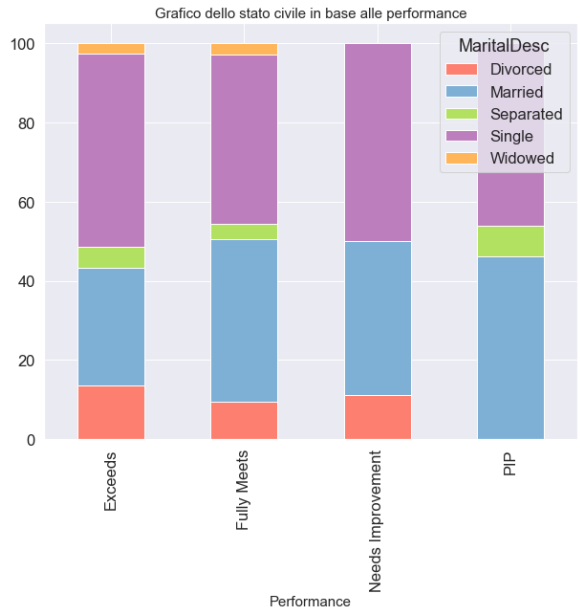


*Figura 9- Livello di performance in base allo stato civile*

Infine, osserviamo come le diverse etnie si suddividono a seconda della performance: il grafico mostra una presenza elevata della categoria *White* nella maggior parte dei diversi livelli di performance. In particolar modo si osserva un gran numero di persone di etnia *Black or African American*, *White* e *Asian* nel livello di performance *Fully Meets* ed un gran numero di persone di etnia *Black or African American* e *White* nei livelli di performance *Exceeds* e *Needs Improvement*. Le persone appartenenti alla categoria *Two or more races* sono invece presenti in piccola parte in ogni categoria di performance.

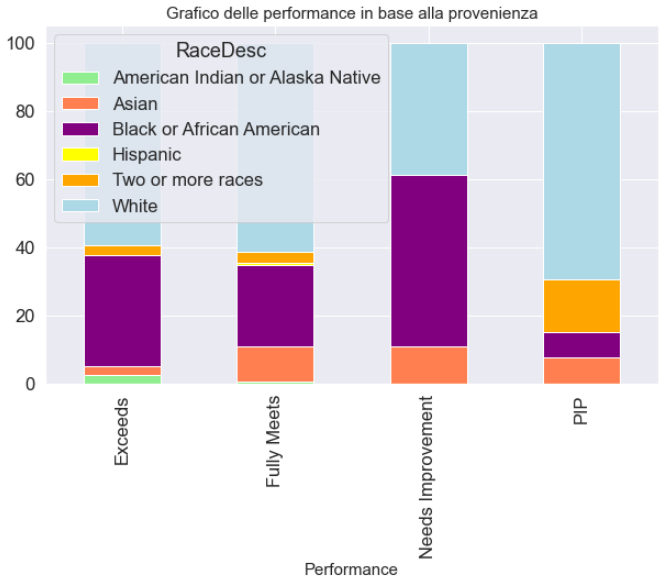


*Figura 10- Livello di performance in base all'etnia*

# 4. Analisi delle Assenze

Un'ulteriore variabile numerica d'interesse è il numero di assenze compiuto da ogni impiegato all'interno dell'azienda. Come già proposto per il salario, visualizziamo prima la distribuzione dei valori della variabile sfruttando un istogramma, per poi concentrarci sul rapporto di quest'ultima con altre features categoriche del dataset.

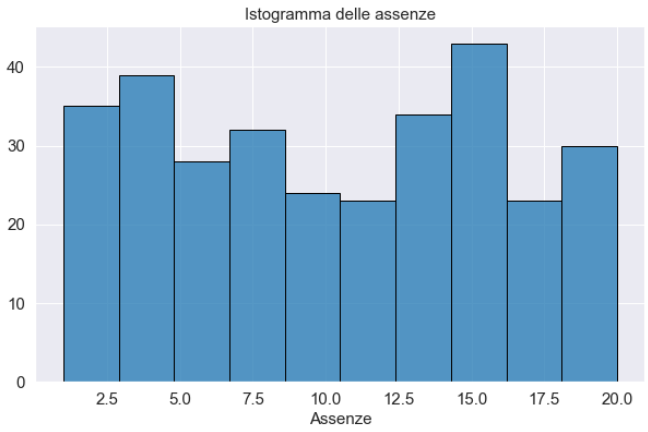


*Figura 11- Frequenza assenze all'interno del Dataset*

Attraverso il grafico risulta evidente un picco maggiore per l'intervallo di assenze tra 15 e 17 ed un picco minore è riscontrabile tra le 2.5 e le 5 assenze. Diventa quindi interessante comprendere quali possano essere le relazioni tra il numero di assenze ed altre categorie del dataset. Ad esempio, il box-plot riportato in basso evidenzia come le assenze si distribuiscono rispetto a livelli diversi di soddisfazione.

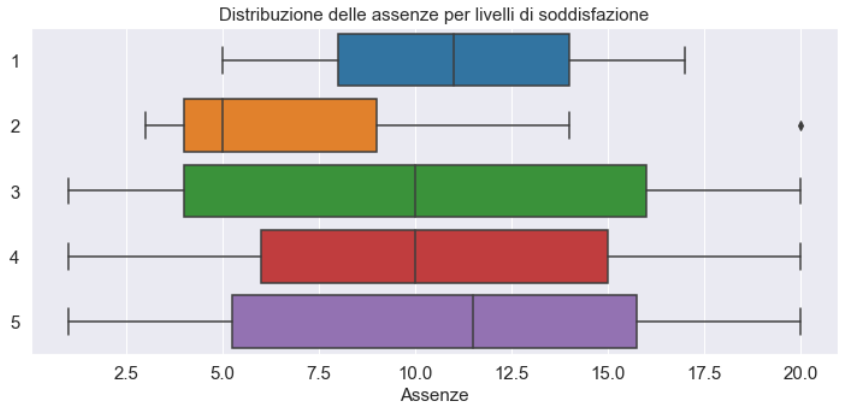


*Figura 12- Distribuzione delle assenze per livelli di soddisfazione*

Interessante è notare come il box-plot per il livello "2" di soddisfazione presenti una mediana sbilanciata verso il primo quartile, con un solo outlier individuato, un impiegato che ricopre la posizione di *Production Technician I*.
Una seconda osservazione interessante può essere fatta considerando i livelli di assenza rispetto allo status degli impiegati. Come è ragionevole pensare, gli impiegati licenziati per una causa precisa presentano un terzo quartile di un valore leggermente maggiore rispetto a quello degli impiegati attivi. Analogamente, il valore del primo quartile per gli impiegati licenziati è discretamente maggiore rispetto al valore per gli impiegati attivi.

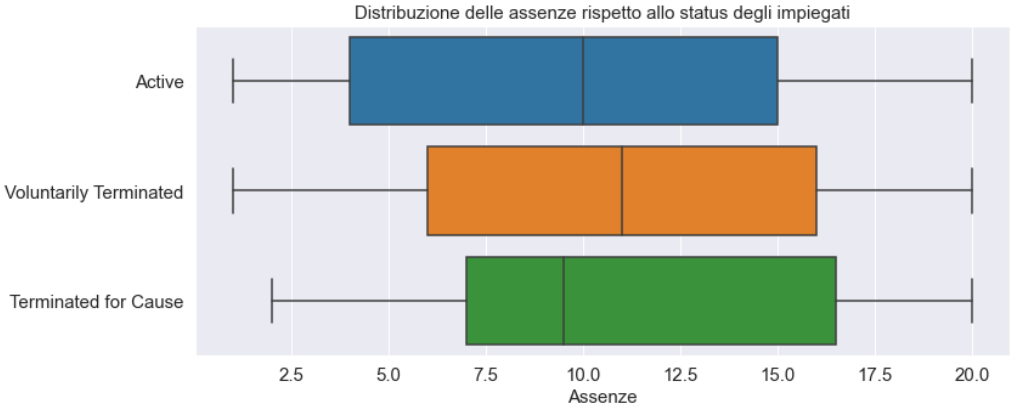


*Figura 13- Distribuzione delle assenze rispetto allo status degli impiegati*

Anche in questo contesto possiamo approfondire l'analisi considerando la media delle assenze compiute da impiegati di diverse categorie: in particolare ci concentriamo sulle assenze compiute dagli impiegati in base allo status di assunzione e alla loro performance sul lavoro.

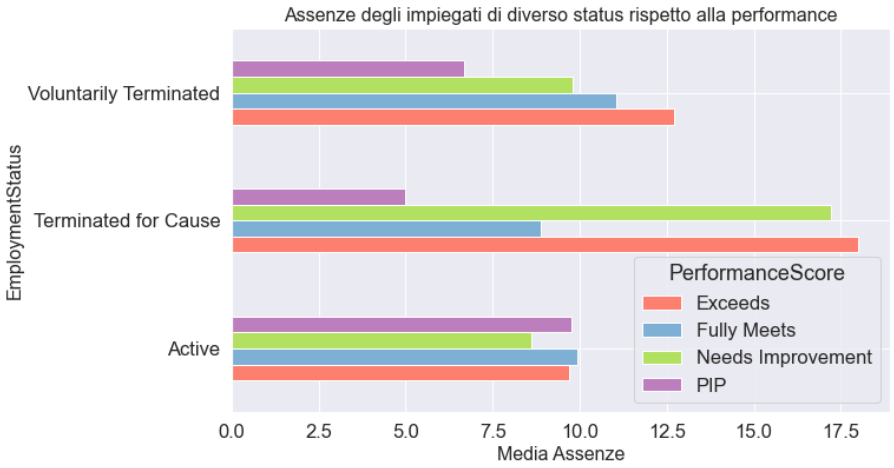


*Figura 14- Assenze degli impiegati di diverso status in base alla performance*

Ci sono alcuni punti degni di nota che emergono dal barplot: è interessante notare come gli impiegati licenziati presentino due picchi nella media delle assenze per le classi *Needs Improvement* ed *Exceeds*, mentre invece gli impiegati che presentano una performance molto bassa e sono stati licenziati presentano una media di assenze particolarmente bassa. Inoltre, c'è un unico impiegato che eccede i livelli di performance ma è stato licenziato per ragioni di "*gross misconduct*", mentre invece gli impiegati con performance sufficienti che hanno subito il licenziamento lo hanno ricevuto in maggioranza per ragioni legate alla presenza sul lavoro.
Infine, si è deciso di osservare come le assenze si suddividessero in base ai diversi dipartimenti presenti in azienda, in base al livello di performance:

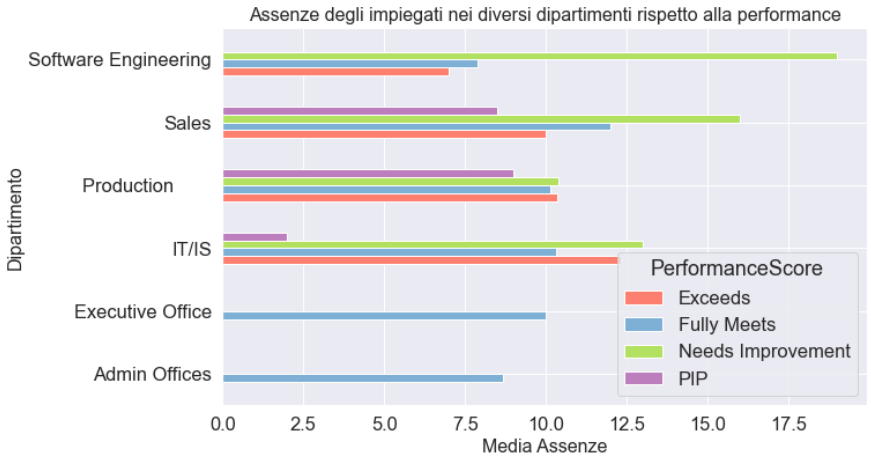


*Figura 15- Assenze degli impiegati nei diversi dipartimenti rispetto alla performance*

Come evidente dal grafico sopra riportato, troviamo un maggior numero di assenze nei dipartimenti di *Software Engineering* e *Sales*, in entrambi i casi le assenze nelle due categorie appartengono al livello di performance *Needs Improvement*. Un elevato numero di assenze è presente anche nella categoria di performance *Exceeds* per quanto

riguarda il dipartimento di *IT/IS*, mostrando come un gran numero di assenze non sia obbligatoriamente correlato ad un calo di performance.

# 5. Correlazione

## 5.1 Correlazione di Pearson

Una volta osservate ed analizzate le features per noi di maggiore interesse, è importante osservare se e in quale modo le features presentano tra loro una correlazione. Per stabilire il rapporto lineare tra coppie di variabili, utilizziamo il coefficiente di correlazione di *Pearson*: la heatmap visualizzata in basso prende in considerazione un sottoinsieme delle variabili originali del dataset, considerando le features numeriche e continue:

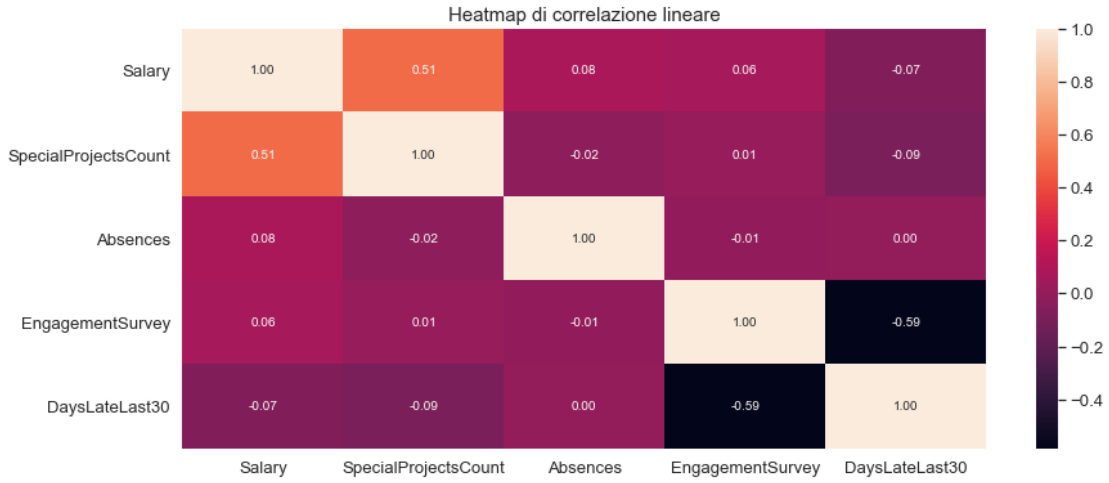


*Figura 16- Heatmap della correlazione di Pearson*

Come visibile dalla mappa la correlazione tra le variabili *Absences* e *Salary* e *Absences* e *SpecialProjectsCount* risulta essere notevolmente bassa, con un indice di correlazione pari a 0.08 e 0.02. L'indice di correlazione positivo più alto è dato dalle variabili *Salary* e *SpecialProjectsCount* con un valore di 0.51, mentre una correlazione negativa interessante è riscontrabile tra *EngagementSurvey* e *DaysLateLast30* con un indice pari a -0.59.

A partire da queste considerazioni, ci concentriamo sul rapporto tra le due coppie di variabili continue aventi un indice di correlazione interessante, ovvero *SpecialProjectCount* e *Salary* e *DaysLateLast30* e *EngagementSurvey*: utilizziamo uno scatterplot per visualizzare la relazione tra le due coppie di variabili.

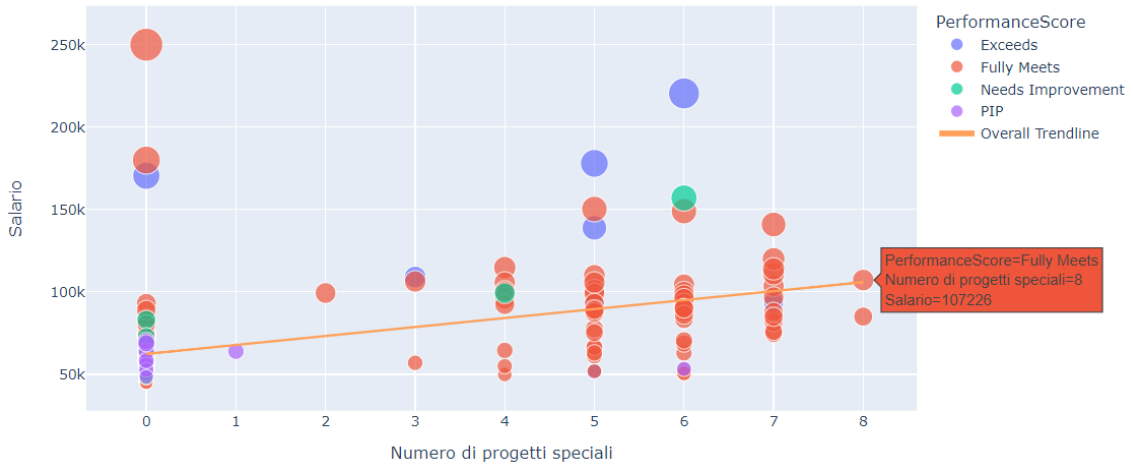


*Figura 17- Scatterplot di SpecialProjectsCount e Salary con distinzione per PerformanceScore*

Risulta evidente che una quantità consistente di impiegati dal salario basso non sono coinvolti in alcun progetto speciale; allo stesso tempo notiamo due impiegati coinvolti in otto progetti speciali ma con un salario relativamente inferiore rispetto a numerosi altri impiegati che hanno partecipato a meno progetti. Evidenziamo in ultimo come la maggioranza degli impiegati che hanno preso parte a due o più progetti speciali presentino un punteggio di performance sufficiente e ci siano solo pochi di questi per i quali è necessario un miglioramento. La retta di regressione costruita all'interno dello scatterplot mostra una tendenza lineare, ma non approssima in modo soddisfacente la loro distribuzione.

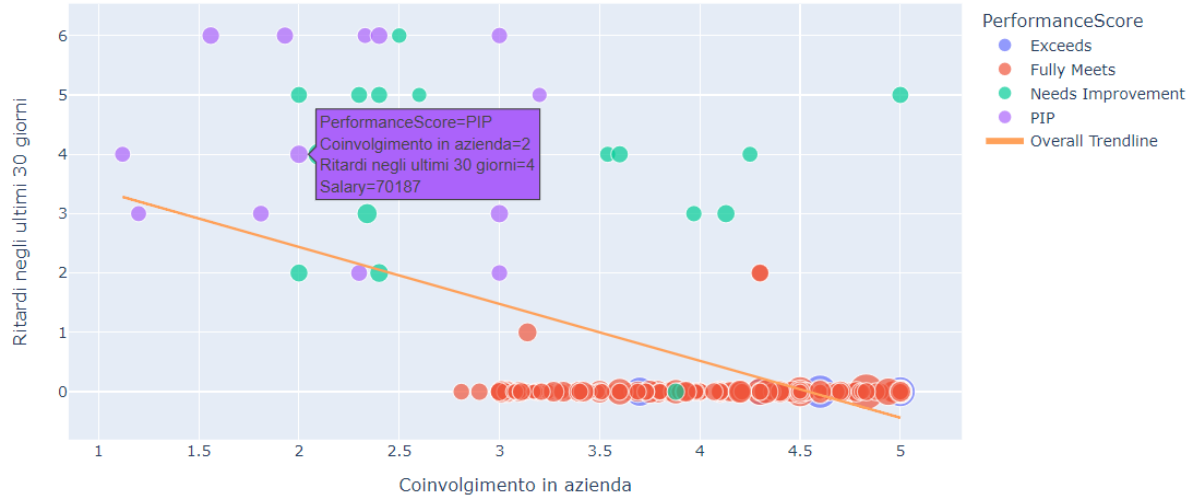


*Figura 18- Scatterplot di EngagementSurvey e DaysLateLast30 con distinzione per PerformanceScore*

Il grafico sopra riportato evidenzia bene la correlazione negativa specificata dal coefficiente di *Pearson* tra la variabile *EngagementSurvey* e *DaysLateLast30*. Il risultato è intuitivamente ragionevole: ci aspettiamo che impiegati più coinvolti

all'interno dell'azienda tendano a non presentarsi in ritardo rispetto ad impiegati meno coinvolti. Lo scatterplot mette in luce anche una relazione interessante con una terza variabile, il punteggio degli impiegati: tutti gli impiegati che hanno fatto ritardo a lavoro tre o più volte negli ultimi 30 giorni presentano punteggi di performance inferiori alla sufficienza, ed è evidente come in particolare gli impiegati con un livello di coinvolgimento minore di 2.0 abbiano il più basso punteggio di performance. Per quanto riguarda la retta di regressione, è evidenziata una relazione di tendenza negativa tra i dati ma, per il modo in cui i valori si distribuiscono nel plot, anche in questo caso la retta non dà un'idea soddisfacente del rapporto negativo tra le due features considerate.

## 5.2 Correlazione di Spearman e Kendall

Come già precedentemente sottolineato, dal momento che le variabili non presentano una normale distribuzione, può essere utile utilizzare test di correlazione non parametrici: utilizziamo dunque il coefficente di correlazione di Spearman e Kendall:


*Figura 19-Heatmap della correlazione di Spearman*

Anche valutando il coefficiente di correlazione di Spearman risultano evidenti la correlazione positiva tra *SpecialProjectsCount* e *Salary* e la correlazione negativa, questa volta con un valore di correlazione leggermente minore rispetto a quanto riportato con Pearson, tra *DaysLateLast30* ed *EngagementSurvey*. Potremmo

aspettarci che le valutazioni introdotte in questo caso siano leggermente più affidabili rispetto a quanto riportato nella heatmap considerando il coefficiente di Pearson, data la tendenziale non-normalità delle variabili e la maggior robustezza del coefficiente di Spearman rispetto agli outliers delle distribuzioni.


*Figura 20- Heatmap della correlazione di Kendall*

Il coefficiente di correlazione di Kendall invece mostra risultati che si discostano fortemente da quelli analizzati precedentemente. In primo luogo, enfatizza una correlazione lievemente negativa tra *SpecialProjectsCount* e *Salary*, contrariamente a quanto riportano le due heatmap sopra riportate. Dall'altro lato, stabilisce correlazioni negative per un numero maggiore di coppie di variabili rispetto a quanto non presenti il coefficiente di Pearson. I risultati, anche in questo caso, enfatizzano una anti-correlazione notevole tra *EngagementSurvey* e *DaysLateLast30* (ci aspetteremmo allora che il ranking di una variabile è quasi completamente opposto al ranking dell'altra).

### 5.3 Correlazione senza Outliers

Osservati i diversi metodi di correlazioni tra le variabili prese in considerazione, può essere interessante vedere se la rimozione degli outliers precedentemente individuati possa impattare sul calcolo della correlazione tra variabili quantitative:

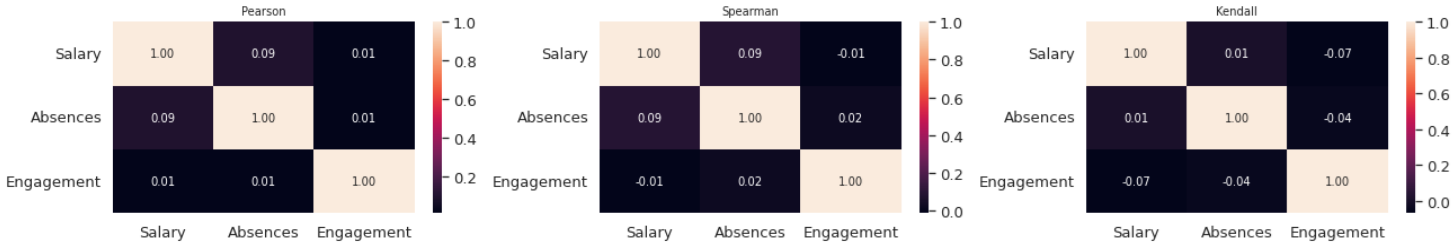


*Figura 21- Heatmap di correlazione senza outliers*

In questo caso, tutti i coefficienti di correlazione utilizzati non enfatizzano alcuna correlazione interessante. Rimuovere gli outliers nullifica le già lievi correlazioni trovate considerando tutte le 311 osservazioni. Specifichiamo che nell' analisi di correlazione i valori di *SpecialProjectsCount* e *DaysLateLast30* non sono riportati: questo perché l'approccio per l'individuazione degli outliers, il range interquartile, ha come risultato estremo quello di identificare come outliers per le due variabili in questione tutti i valori che non corrispondono a 0, ciò a conseguenza del fatto che in entrambi i casi il primo e terzo quartile hanno valore pari a 0.

## 6. Riferimenti

-Dataset: Human Resources Data Set,
https://www.kaggle.com/datasets/rhuebner/human-resources-data-set