ANALISI DI UN DATASET DI RISORSE UMANE

Martina Leocata
636189

Irene Mondella
584285

Elena Scaglione
645638

1. INTRODUZIONE

Il presente report si propone di riassumere l'analisi da noi effettuata su un dataset di risorse umane, durante la quale abbiamo esplorato la natura dei dati e la loro distribuzione, e abbiamo cercato eventuali correlazioni tra di essi. Cercheremo di spiegare le modalità e le ragioni del nostro operato, e di mostrare quali risultati sono emersi da questa indagine.

2. DATA UNDERSTANDING

Il dataset fornito riporta i dati delle risorse umane di un'azienda: informazioni anagrafiche (quali nome, data di nascita, genere, etnia e stato civile), posizione lavorativa, salario, status del dipendente, valutazione della sua performance, giorni di ritardo e di assenza, nonché i risultati dei sondaggi di soddisfazione e coinvolgimento all'interno dell'azienda.

Il dataset è composto da 311 osservazioni (ognuna corrispondente ad una riga) e 36 variabili, distinguibili in qualitative (categoriche e ordinali) e quantitative (discrete e continue). Alle colonne originarie, abbiamo aggiunto la variabile quantitativa e continua 'Age' (età), ricavabile a partire dall'anno di nascita assumendo che i dati siano stati estratti nell'anno corrente, insieme ad altre variabili utili per ulteriori analisi, in particolare quelle riguardanti le assunzioni (ad esempio la variabile 'Hiring_Year') e quelle sul salario (come 'Salary_zscore').

In questo modo, il dataset ottenuto è composto da 41 variabili, delle quali soltanto 7 quantitative continue nel dataset originario: 'Salary' (salario), 'PerfScoreID' (punteggio della performance lavorativa), 'EngagementSurvey' (risultato del sondaggio di coinvolgimento nell'azienda), 'EmpSatisfaction' (indice di soddisfazione del dipendente nell'azienda), 'SpecialProjectCount' (numero di progetti speciali effettuati dal dipendente), 'DaysLateLast30' (giorni di ritardo nell'ultimo mese), 'Absences' (assenze) .

Abbiamo quindi approfondito l'analisi di queste variabili insieme a quella relativa all'età, andando a calcolare i vari indici statistici principali come media, deviazione standard, valore minimo, valore massimo, primo percentile, secondo percentile (cioè la mediana) e terzo percentile, riportati in Tabella 1. Analizzando la variabile 'Age', notiamo che il valore della media e della mediana sono piuttosto vicini, rispettivamente 43,41 e 42. Anche il rapporto tra deviazione standard e media risulta essere piuttosto basso, intorno al 20%, segno che non c'è troppa varianza nella distribuzione e che la media è tutto sommato affidabile. Andando a visualizzare l'istogramma, però, notiamo come, sebbene media e mediana si trovino nell'intervallo tra i 40 e i 45 anni, l'intervallo di età più denso risulta essere quello compreso tra i 35 e i 40 anni, come dimostra l'area maggiore del rispettivo bin in Figura 1.

	Engagement Survey	**Salary**	**Emp Satisfaction**	**Special Projects**	**Perf ScoreID**	**DaysLate Last30**	**Age**	**Absences**
count	311	311	311	311	311	311	311	311
mean	4,11	69020,68	3,89	1,22	2,98	0,41	43,41	10,24
std	0,79	25156,64	0,91	2,35	0,59	1,29	8,87	5,85
min	1,12	45046	1	0	1	0	30	1
25%	3,69	55501,5	3	0	3	0	36	5
50%	4,28	62810	4	0	3	0	42	10
75%	4,7	72036	5	0	3	0	49	15
max	5	250000	5	8	4	6	71	20

Tabella 1 Indici statistici delle variabili quantitative di interesse

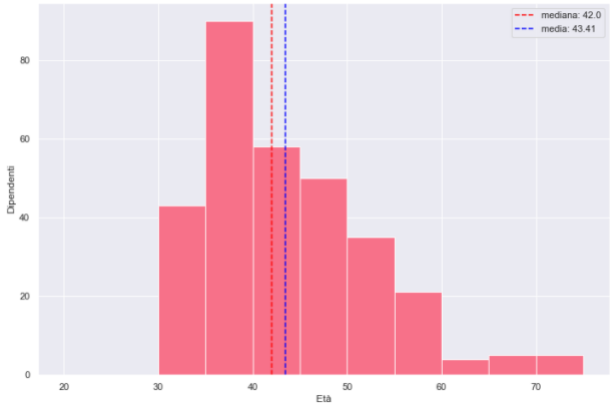

Figura 1 Distribuzione età dei dipendenti

Per quanto riguarda la variabile 'Salary', la deviazione standard è particolarmente elevata, con un rapporto con la media pari a circa il 36%, e ciò indica che i valori si discostano dalla media in modo non trascurabile. Questo è confermato anche dalla differenza tra media e mediana, fenomeno che si verifica quando i dati presentano un numero considerevole di valori anomali (*outliers*), i quali influenzano molto la media e poco mediana: in questo caso, tutti gli outliers, sia quelli estremi che quelli moderati, presentano valori di molto al di sopra del cinquantesimo percentile, comportando un valore della media superiore rispetto a quello della mediana. Per quanto riguarda il coefficiente di Skewness, infine, il salario presenta un valore di 3,306, il quale conferma che la distribuzione non è normale ma fortemente asimmetrica e schiacciata verso sinistra: c'è una gran quantità di dipendenti che guadagna salari in un intervallo tra i 50k e i 70k, con un picco tra i 60 e i 65 mila dollari, e un esiguo numero di dipendenti che riceve un salario nettamente più alto rispetto alla maggioranza. Perciò, si può dire che la variabile 'Salary' ha una distribuzione di Zipf (*power law distribution*), chiaramente visibile in Figura 2.

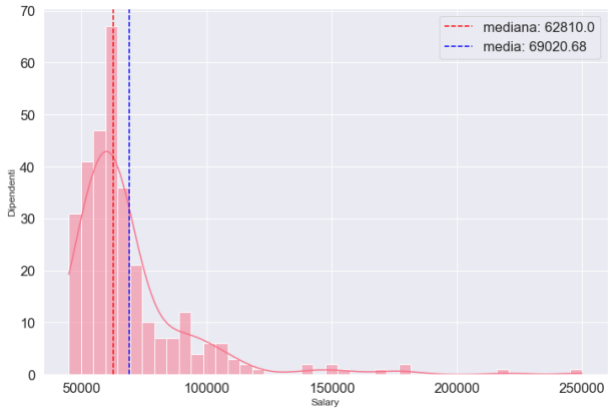

Figura 2 Distribuzione del salario

La distribuzione di Zipf è molto diffusa, ma la presenza di numerosi *outliers* che caratterizza questo tipo di curva può inficiare negativamente sulla ricerca e sulla visualizzazione di eventuali correlazioni tra i dati. Per questo motivo, abbiamo creato un ulteriore dataset escludendo i valori anomali del salario. Così facendo, la distribuzione ottenuta è quella in Figura 3, che appare immediatamente più leggibile della precedente.

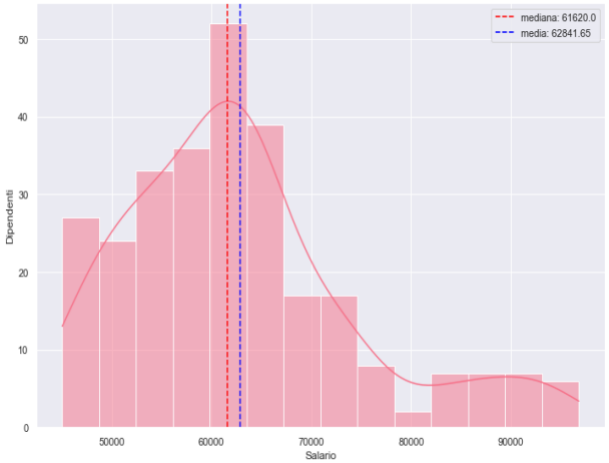

Figura 3 Distribuzione del salario senza *outliers*

Un altro modo in cui si può maneggiare dei dati troppo *skewed* è quello di normalizzarli tramite una trasformazione logaritmica, allo scopo di rendere la distribuzione meno sbilanciata e più interpretabile: abbiamo ricavato il logaritmo dei dati relativi al salario, memorizzandoli in una colonna apposita 'log_salary', abbiamo quindi ottenuto l'istogramma. In questo modo, la curva è più simile a una normale, come è visibile in Figura 4.

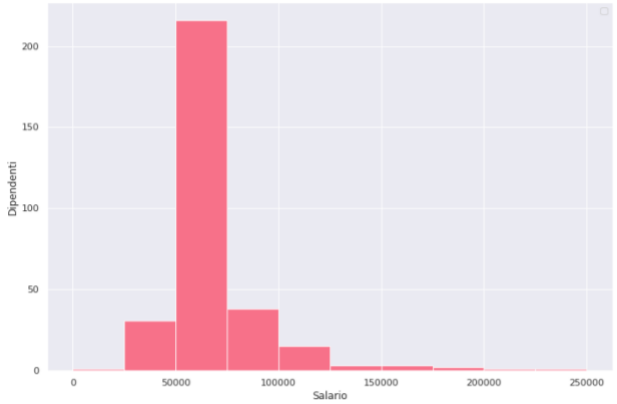

Figura 4 Distribuzione logaritmica del salario

Dopo aver così analizzato la natura del dataset, e aver compreso la tipologia e la distribuzione delle principali variabili, abbiamo provato a individuare eventuali correlazioni tra le feature quantitative. Per fare ciò, abbiamo creato un sottoinsieme del dataset originario, contenente soltanto le variabili numeriche (tra le quali il salario contenente anche gli *outliers*). Abbiamo poi realizzato una matrice di correlazione (*heatmap*), utilizzando il coefficiente di Pearson. L'*heatmap* evidenzia una leggera correlazione positiva tra:

- Coinvolgimento e performance;

- Salario e numero di progetti speciali.

Emerge, invece, una correlazione negativa tra:

- Ritardi nell'ultimo mese e coinvolgimento;
- Ritardi nell'ultimo mese e performance.

Tuttavia, consapevoli del fatto che alcune variabili presentano molti valori anomali, abbiamo deciso di utilizzare anche il coefficiente di Spearmann, poiché è meno sensibile agli *outliers*. La nuova *heatmap* che ne è risultata (Figura 5) presenta le stesse correlazioni già osservate nella precedente, seppur meno forti (sia quelle positive che quelle negative).

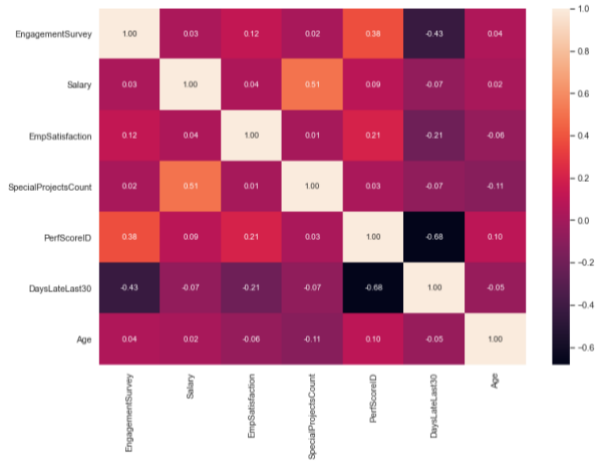

Figura 5 Matrice di correlazione calcolata con il coefficiente di Spearmann

3. PULIZIA DATASET

Uno step importante effettuato in fase di preparazione all'analisi è stato quello del *data cleaning*, che consiste nell'andare a verificare l'esistenza di particolari inconsistenze nel dataset, ad esempio la presenza di *missing values*. Nel nostro caso, il dataset risultava abbastanza pulito, non presentando valori NaN se non nelle variabili 'DateOfTermination', che indica la data dell'interruzione del rapporto lavorativo, e 'ManagerID', ovvero il numero di matricola associato ad ogni manager. In entrambi i casi siamo andati a vedere quali fossero le righe in cui si verificava questa inconsistenza: nel primo i valori NaN erano registrati nei casi di dipendenti il quale 'EmploymentStatus' risultava come attivo, segno del fatto che non si tratta di veri e propri *missing values*; nel secondo si sono registrate invece delle inconsistenze in concomitanza di un determinato valore del 'ManagerName', inconsistenza risolta andando a riempire i valori NaN con la matricola associata a quel particolare manager attestata nelle altre celle.

4. DATA EXPLORATION

L'esplorazione del dataset si è concentrata su specifiche variabili, al fine di indagare tre principali macroaree:

1. Com'è composto il corpo dei dipendenti attivi?
2. Come si distribuisce il salario? Esistono disparità? In quali reparti si guadagna di più?
3. Qual è la performance dei lavoratori? Cambia il base al reparto? C'è una correlazione tra la performance e il grado di coinvolgimento e di soddisfazione dei dipendenti?

4.1 Composizione dei dipendenti attivi

Nel dataset sono presenti 311 lavoratori assunti tra il 2006 e il 2018: 207 attualmente in servizio, 104 tra dimessi e licenziati.

L'età media dei lavoratori attivi è di circa 42,77 anni, con una mediana che si discosta di poco (41 anni); l'intervallo più popolato è quello nella fascia tra 35 e 40 anni.

Per quanto riguarda il genere, c'è una leggera prevalenza di donne (65,9%) rispetto agli uomini (34,1%).

Gli impiegati sono per la maggioranza bianchi (59,9%), seguiti da neri o afroamericani (24,6%), asiatici (9,7%) e altre etnie (5,8%).

La stragrande maggioranza dei dipendenti, il 61%, appartiene al settore della produzione (126 dipendenti), seguiti dal 19% del settore IT/IS, (40 dipendenti), 13% delle vendite (26 dipendenti) e dal 3% del settore del software engineering e degli uffici amministrativi (7 dipendenti ciascuno) e, infine, l'ufficio esecutivo, con un solo dipendente, ovvero la CEO.

4.2 Salario

Come già rilevato, la distribuzione del salario segue la legge di potenza. Questo si traduce in una maggioranza di dipendenti attivi, più del 70%, che guadagna annualmente dai 45K ai 70K e una esigua minoranza che guadagna fino ai 250K. Dall'analisi del salario in base al genere, non emergono particolari disparità (Figura 6): la mediana si attesta intorno ai 64K per gli uomini e ai 63K per le donne; il terzo quartile dello stipendio degli uomini risulta leggermente più ampio, con una differenza di circa 4K in più rispetto alle donne.

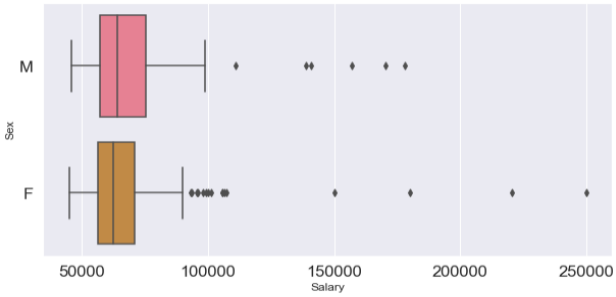

Figura 6 Boxplot del salario dei dipendenti attivi in base al genere

Esiste infatti un reparto, quello delle vendite, in cui gli uomini guadagnano più delle donne, ma senza differenze statisticamente significative. Filtrando gli impiegati che guadagnano sopra i 100K, 10 sono donne e 6 uomini: la posizione di CEO e quella di CIO sono occupate da due donne; anche la terza posizione più remunerativa, direttrice delle vendite, è occupata da una donna (Tabella 2). Questi sono infatti gli outliers più estremi visibili nel boxplot.

	Nome	Posizione	Salario	Z score
1	King, Janet	President & CEO	250000	7.21
2	Zamora, Jennifer	CIO	220450	6.03
3	Houlihan, Debra	Director of Sales	180000	4.42
4	Foss, Jason	IT Director	178000	4.34
5	Corleone, Vito	Director of Operations	170500	4.04
6	Monroe, Peter	IT Manager - Infra	157000	3.50
7	Roper, Katie	Data Architect	150290	3.24
8	Roup,Simon	IT Manager - DB	140920	2.86
9	Dougall, Eric	IT Manager - Support	138888	2.78
10	Champaigne, Brian	BI Director	110929	1.67

Tabella 2 I 10 dipendenti che guadagnano di più

Anche dal punto di vista dell'etnia non emergono disparità salariali (Figura 7): le mediane del salario percepito dai diversi gruppi non differiscono molto (non considerando l'unico impiegato ispanico). In base alla mediana, i dipendenti che guadagnano di più sono i nativi americani (che però sono solo 3), i neri o afroamericani (51 dipendenti) e gli asiatici (20 dipendenti); seguono poi i bianchi (124 dipendenti) e chi ha un'origine multietnica (8 dipendenti), con una mediana che differisce di circa 8K dalla mediana del salario dei nativi americani. I dipendenti che guadagnano più di 100K sono neri o afroamericani (8), bianchi (5, tra cui le due CEO e CIO) e asiatici (3) (Tabella 3).

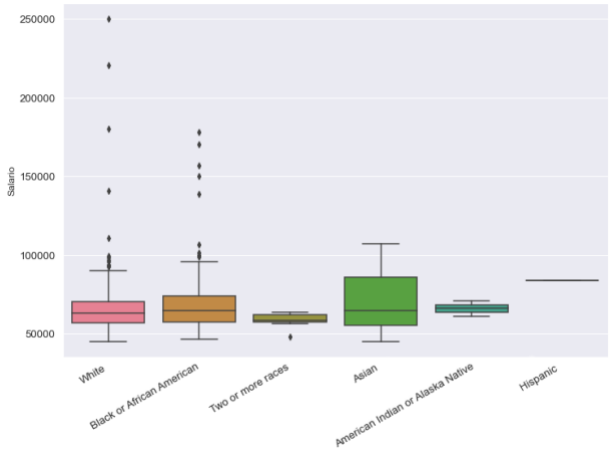

Figura 7 Boxplot del salario dei dipendenti attivi in base all'etnia

Etnia	Mediana salario
Hispanic	83667.0
American Indian or Alaska Native	66149.0
Black Or African American	64816.0
Asian	64355.5
White	62933.5
Two Or More Races	58284.0

Tabella 3 Mediana salario in base all'etnia

Sono stati analizzati gli stipendi in base al reparto: l'ufficio esecutivo è naturalmente quello con il salario più alto (250K), seguito dal reparto di software engineering (mediana di 93K) e IT/IS (mediana di 90K). I reparti dove si guadagna di meno sono quello di produzione (mediana di 60K) e vendite (mediana di 64K).
Data la correlazione rilevata tra salario e progetti speciali, abbiamo deciso di rappresentare le due variabili in un boxplot, per indagare meglio il fenomeno (Figura 8) considerando in questo caso impiegati attivi e inattivi.

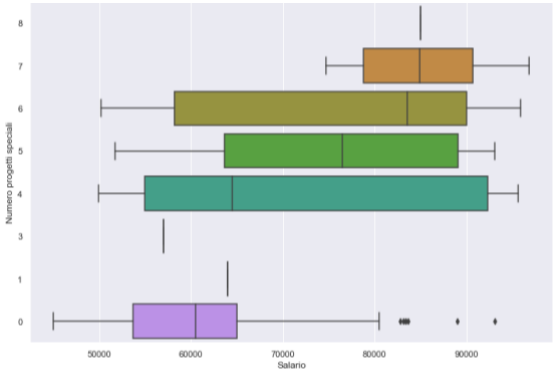

Figura 8 Boxplot del salario senza outliers in base al numero di progetti speciali

Notiamo infatti che la mediana del salario aumenta proporzionalmente al numero di progetti. Tuttavia, osservando la distribuzione dei progetti speciali per reparto, vediamo anche che la maggior parte di essi appartengono al reparto IT/IS (246 progetti) e software engineering (46 progetti): questi sono anche i reparti dove la mediana del salario è più alta. Possiamo concludere quindi che, molto probabilmente, i progetti speciali risultano correlati al salario in quanto vengono portati avanti dai reparti più remunerativi in termini di salario.

4.3 Assunzioni e interruzioni del rapporto lavorativo

Un ulteriore aspetto indagato riguarda il numero di dipendenti attivi e inattivi.
Abbiamo analizzato l'andamento delle assunzioni e delle interruzioni del rapporto lavorativo, che comprendono sia le dimissioni che i licenziamenti. Raggruppando i dipendenti per il loro 'EmploymentStatus', emerge che 2/3 dei dipendenti

riportati nel dataset risultano ancora attivi, mentre, tra i lavoratori che hanno interrotto il loro rapporto con l'azienda, la maggior parte ha dato volontariamente le dimissioni.
Basandoci sulla colonna da noi creata 'Hiring_Year', emerge, per quanto riguarda l'andamento degli assunti, una distribuzione bimodale, con un picco nel 2011 e uno nel 2014 (Figura 9). Per quanto riguarda le differenze di genere, le assunzioni di donne sono più numerose rispetto agli uomini, in particolare tra il 2010 e il 2011. Questa tendenza è confermata misurando il valore della media: il numero medio di donne assunte risulta quasi il triplo rispetto a quello degli uomini.

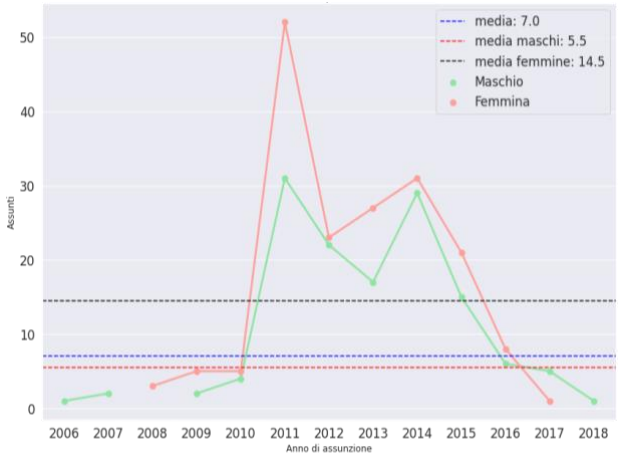

Figura 9 Lineplot dei dipendenti assunti dal 2006 al 2018, divisi per genere

Per quanto riguarda l'interruzione del rapporto lavorativo, ci siamo basati sulla variabile 'TerminationYear', notando come la stragrande maggioranza degli ex dipendenti si sia dimessa piuttosto che licenziata, pari a circa l'85% del totale degli ex dipendenti. Il periodo in cui si registra la maggior parte delle interruzioni va dal 2014 e il 2016. I licenziamenti invece si registrano soprattutto tra il 2013 e il 2018, con un picco nel 2015.
Indagando sui motivi dell'interruzione, infine, notiamo come la maggior parte dei dipendenti dimessi hanno interrotto la loro relazione lavorativa per cambiare posizione, perché infelici o perché desiderosi di un salario maggiore; i licenziati, invece, sono stati allontanati per motivi legati alla frequenza, perché non si sono presentati più a lavoro o per motivi di scarsa performance.

4.4 Performance e gradimento

Abbiamo infine esplorato il rapporto tra performance e gradimento dei dipendenti, espresso in termini di soddisfazione (colonna 'EmpSatisfaction') e di coinvolgimento (colonna 'EngagementSurvey'), osservando anche come queste variabili si distribuiscono lungo i singoli dipartimenti.

Un primo aspetto che notiamo è la correlazione tra le performance e coinvolgimento, come mostrato in Figura 10.

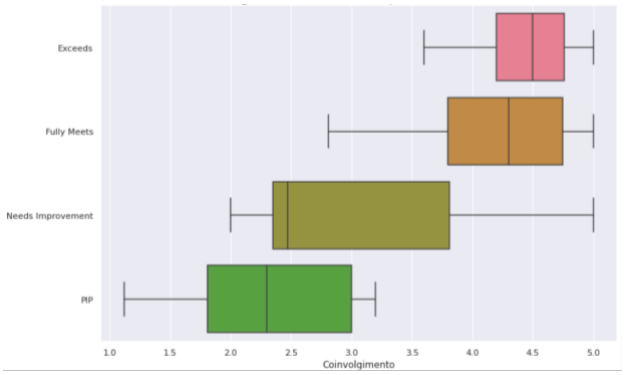

Figura 10 Boxplot del coinvolgimento in base alla performance

Come si vede, al migliorare delle prestazioni del dipendente, risulta maggiore anche il tasso di coinvolgimento dello stesso: dato il valore massimo del coinvolgimento a 5, i valori più alti si registrano tra i dipendenti le cui prestazioni sono etichettate come eccellenti ('Exceeds'), mentre quelli più bassi sono associati ai dipendenti con le prestazioni peggiori ('PIP').
Il passo successivo è stato dunque quello di andare a vedere come le due variabili si distribuiscono lungo i reparti, allo scopo di indagare sull'eventuale presenza di differenze significative. Abbiamo perciò, in un caso, riorganizzato la colonna del dataset relativa al dipartimento ('Department') in base alla variabile associata alla soddisfazione ('EmpSatisfaction'), memorizzandola in un nuovo dataframe; abbiamo quindi fatto lo stesso per il dato della performance ('PerformanceScore'). In questo modo abbiamo ottenuto la distribuzione del numero di dipendenti di ogni reparto per performance, in un caso, e in soddisfazione, nell'altro. Un'ulteriore modifica che abbiamo apportato è stata normalizzare i valori ottenuti in percentuale, in modo tale da eliminare le differenze tra reparti dovute alle differenze di numerosità.
A questo punto abbiamo deciso di utilizzare un bar chart in versione stacked, in maniera tale da poter visualizzare a meglio le differenze di distribuzione tra i reparti. I risultati ottenuti sono in Figura 11 e Figura 12:

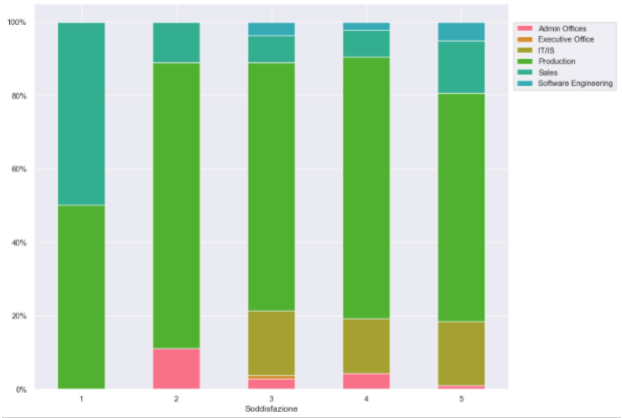

Figura 11 Bar chart stacked di soddisfazione per reparto

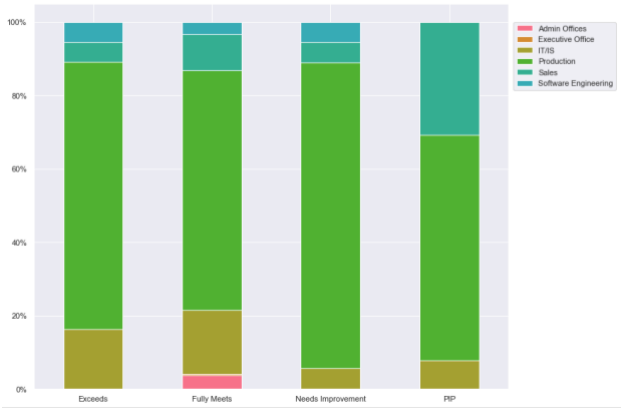

Figura 12 Bar chart stacked di performance per reparto

Dai grafici ottenuti notiamo delle tendenze simili: il reparto caratterizzato dalla performance migliore e dal coinvolgimento maggiore risulta essere quello dell'IT/IS, insieme a quello di software engineering, per quanto riguarda la soddisfazione; quello in cui sia la performance che il gradimento sono invece peggiori risulta essere quello delle vendite: è questo, infatti, quello in cui si registra il più alto tasso di 'PIP' e il più basso di 'Exceeds' e, in maniera ancor più netta, quello a cui appartiene il 50% dei dipendenti la cui soddisfazione è pari a 1.

5. CONCLUSIONI

In base alle analisi da noi condotte, possiamo dire che l'azienda da noi analizzata mostra un corpo dipendenti eterogeneo dal punto di vista etnico, con una leggera prevalenza di donne rispetto agli uomini e tendenzialmente giovane.

Il salario, prevedibilmente, dimostra una distribuzione non gaussiana, segno che i dipendenti ai vertici di essa percepiscono un salario nettamente superiore a quello della stragrande maggioranza dei dipendenti. Tuttavia, analizzando la distribuzione lungo le variabili anagrafiche più delicate, l'azienda si dimostra equa, non essendoci disparità di distribuzione a livello di genere o etnia.

Per quanto riguarda performance e coinvolgimento, infine, le due variabili risultano correlate: infatti, all'aumentare della performance aumenta anche il coinvolgimento. Il reparto migliorabile, inoltre, risulta essere quello delle vendite, registrando valori bassi in entrambi i casi.