Giacomo Fidone, Vittorio Mussin, Virginia Ranciaro, Ilaria Salogni

Progetto di gruppo per il modulo 'Analisi di dati' – Relazione finale

Questa relazione si propone di illustrare l'iter di svolgimento del progetto da noi realizzato per l'esame del modulo Analisi di Dati.

Introduzione

L'obiettivo del progetto è l'analisi del dataset 'Human Resources Data Set – Version 14', un dataset sintetico contenente i dati dei dipendenti di un'azienda americana fittizia.

L'analisi è stata svolta combinando operazioni di statistica descrittiva con strumenti di *data visualization*: ogni indagine è rappresentata visivamente mediante la generazione di grafici che rendono più immediata la lettura e la comprensione dell'analisi.

Il progetto è stato elaborato su Jupyter Notebook, in forma di documento testuale interattivo contenente celle di codice sorgente eseguibile, funzionale alla creazione dei grafici e all'esecuzione dei calcoli statistici. Il linguaggio di programmazione utilizzato per la scrittura del codice è Python, mentre gli elementi testuali del documento sono stati scritti in linguaggio Markdown.

Il dataset HR si compone di 36 variabili e 311 osservazioni; le variabili sono di tipo sia quantitativo che qualitativo, e contengono informazioni di vario genere relative a ciascun dipendente. Si è convenzionalmente deciso di suddividerle in macrogruppi tematici. Un dizionario completo delle variabili è disponibile a questo indirizzo: https://rpubs.com/rhuebner/hrd_cb_v14.

1. Data cleaning e Data enrichment

La prima fase dell'analisi svolta consiste in una serie di operazioni di pulizia e arricchimento dei dati che sono state effettuate direttamente sul dataset di partenza.

1.1. Correzione del dataset

Nell'esplorazione del dataset abbiamo riscontrato alcuni errori di registrazione dei dati, che abbiamo opportunamente corretto prima di procedere con l'analisi (cfr. Jupyter Notebook per la descrizione dei casi specifici).

1.2. Enrichment del dataset

Si è scelto di inserire nel dataset delle nuove variabili, non presenti nel dataframe di partenza, funzionali alla conduzione della nostra analisi:

1. *Age* (quantitativa discreta), che indica l'età del dipendente;
2. *LateOrPunctual* (nominale), che indica se il dipendente ha effettuato almeno un ritardo negli ultimi 30 giorni oppure è stato puntuale (rielaborazione della variabile *DaysLateLast30*);
3. *DurataContratto* (quantitativa discreta), che indica la durata in giorni del contratto dei dipendenti non più attivi;

4. *MotiviLicenziati* e *MotiviDimessi* (nominali), ciascuna delle quali raggruppa i motivi di licenziamento e di dimissioni (espressi nella variabile *TermReason*) in due macrocategorie:
    - *MotiviLicenziati*: “performance” o “condotta”;
    - *MotiviDimessi*: “insoddisfatti” o “scelte personali”.
5. *Races* (nominale), costruita sostituendo con “Hispanic” i valori di *RaceDesc* dei dipendenti che nella colonna *HispanicLatino* presentavano la modalità “yes”.

1.3. Gestione degli outliers

Dopo la pulizia e l’arricchimento del dataset, abbiamo definito un insieme di funzioni da invocare, in caso l’analisi lo richieda, per il controllo e la gestione degli outliers nelle variabili quantitative del dataset. Abbiamo definito due funzioni, **upperfence()** e **lowerfence()**, per calcolare il limite massimo e minimo oltre il quale i valori di una distribuzione vengono considerati outliers. Abbiamo usato le seguenti formule, basate sul valore dello scarto interquartile (IQR):

- Limite superiore = Q3 + (1.5 * IQR)
- Limite inferiore = Q1 – (1.5 * IQR)

Usando la funzioni lowerfence() e upperfence() è stata poi definita una terza funzione **check_outliers()** per verificare l’eventuale presenza di outliers in una distribuzione. Poiché la definizione di outliers è basata sullo scarto interquartile, la funzione verifica anche che l'IQR della distribuzione non sia nullo.

Abbiamo successivamente definito una funzione **df_no_outliers()** che, dati un dataframe ed una sua colonna, restituisce una copia del dataframe in cui la colonna è stata ripulita dagli outliers – se la colonna di partenza non ha outliers, la funzione restituisce il dataframe originale.

L’ultima funzione definita è **df_only_outliers()**: dati un dataframe ed una sua colonna, restituisce una copia del dataframe che contiene soltanto gli outliers – questa funzione, sebbene non abbia trovato applicazione nel nostro progetto, ci è sembrata utile da definire per i casi in cui ci sia necessità di isolare ed analizzare i valori anomali.

2. Test di normalità e di affidabilità della media

Insieme alle funzioni per la pulizia e l’arricchimento del dataset, si è deciso di definire due funzioni utili a controllare la normalità di una distribuzione a verificare l’affidabilità dei valori medi.

La prima funzione per il controllo della normalità, denominata **is_normal()**, verifica che la *skewness* della distribuzione sia compresa tra -0.5 e 0.5. Per il valore del coefficiente di curtosi, invece, abbiamo deciso di considerare come range di normalità un intervallo compreso tra -0.3 e 0.3; in caso di esito positivo, il valore della curtosi potrà essere ulteriormente indagato.

La funzione **mean_reliable()**, invece, verifica l’affidabilità della media di una distribuzione prendendo in considerazione il valore della deviazione standard. La media viene considerata un indice di centralità attendibile laddove la deviazione standard è strettamente inferiore al 30% della media stessa. Laddove possibile, è stata usata la media per descrivere una distribuzione – in caso di distribuzioni non normali, si è scelto di utilizzare il valore mediano perché più robusto rispetto agli outliers.

3. Performance dei dipendenti

In questa sezione abbiamo analizzato la performance dei dipendenti – espressa dalla variabile nominale *PerformanceScore* e dalla variabile numerica *PerfScoreID*, considerando i possibili fattori di correlazione.

3.1. Distribuzione delle variabili di performance

È stata innanzitutto verificata l'affidabilità della media per la distribuzione della variabile *PerfScoreID* e sono stati calcolati gli indici di centralità (cfr. Jupiter Notebook).

Si è scelto di rappresentare la distribuzione globale dei punteggi di performance all'interno dell'azienda con un grafico a torta [1], dal quale emerge che solo il 10% dei dipendenti richiedono un intervento (*Needs Improvement*) o rientrano già in un 'Performance Improvement Plan' (*PIP*).

L'analisi della performance nelle varie sezioni aziendali ha restituito invece delle variazioni, seppur minime; la sezione *Sales* è risultata l'unica con una performance inferiore alla media globale (2.77). Un solo valore coincide con la media globale (*Production*), mentre i restanti si collocano sopra la media globale, con un picco nella sezione di *Software Engineering*.

La rappresentazione della distribuzione in percentuale dei punteggi di performance all'interno delle diverse sezioni è stata effettuata mediante un grafico a barre *stacked* [2]. Dal grafico è emerso che le sezioni *Sales* e *Production* sono, nell'ordine, le sezioni dove sono impiegati più dipendenti con un basso punteggio di performance.

3.2. Fattori di correlazione

Abbiamo in seguito indagato i fattori che potrebbero incidere sul punteggio di performance dei dipendenti. Per la valutazione delle possibili correlazioni tra le variabili quantitative del dataset, abbiamo prima verificato la normalità di ciascuna distribuzione. Dai test è emerso che nessuna variabile quantitativa del dataset presenta una distribuzione normale. Avendo deciso di rappresentare le correlazioni della variabile *PerfScoreID* mediante una heatmap [3], abbiamo quindi utilizzato il coefficiente di Spearman (ρ) e non il coefficiente di default (Pearson R), in quanto quest'ultimo è applicabile soltanto a distribuzioni di tipo normale.

La heatmap [3] restituisce un quadro abbastanza completo delle possibili correlazioni. Sono emersi alcuni valori significativi:

- Una correlazione debolmente positiva (0.39) tra la performance e l'engagement.
- Una correlazione negativa (-0.68) tra la performance e i giorni di ritardo (negli ultimi 30 giorni) – curiosamente, non si registra una correlazione analoga con le assenze (*Absences*).
- Una correlazione debolmente negativa (-0.43) tra i giorni di ritardo e l'engagement.

Un altro valore che potrebbe avere una correlazione con la performance è lo stipendio (*Salary*). Il coefficiente di correlazione presentato dalla heatmap (0.08), di per sé poco significativo, non è attendibile, perché in questo caso lo stipendio deve essere considerato non globalmente, ma all'interno di una singola posizione. L'analisi della correlazione stipendio-performance è stata quindi svolta soltanto per posizione.

3.3. Correlazione con i ritardi a lavoro

Come già detto, dalla heatmap abbiamo riscontrato una netta correlazione negativa tra la performance del dipendente ed il numero di ritardi sul posto di lavoro. Abbiamo quindi voluto verificare se la distribuzione dei ritardi nei dipartimenti, mostrata nel grafico a barre [4], sia correlata alla distribuzione della performance. È effettivamente emerso che le sezioni aziendali in cui si riscontrano quote più alte di dipendenti ritardatari sono le stesse che presentano una performance media più bassa e quote percentuali maggiori di dipendenti con basso indice di performance, cioè *Sales* e *Production*. Questi ultimi sono, in base al calcolo dei range per sezione, anche i dipartimenti con dipendenti che hanno effettuato più ritardi (fino a 5 e 6 ritardi per dipendente, contro i 4 massimi delle altre sezioni).

Vista la correlazione negativa tra ritardi ed engagement osservata nell'heatmap [3], si è deciso di analizzare l'andamento della variabile *EngagementSurvey* sia a livello globale che in ciascuna sezione aziendale. Il livello medio di engagement di tutti i dipendenti, con una media di 4.11 e una deviazione standard di 0.79, è da considerarsi positivo.

Nel grafico a barre [5] abbiamo messo in correlazione il livello di engagement dei dipendenti e i giorni di ritardo effettuati, all'interno di ciascuna sezione aziendale: come mostrato, i dipendenti che effettuano un maggior numero di ritardi sono anche quelli con livelli di engagement più bassi in quasi tutte le sezioni aziendali – l'unica eccezione è rappresentata dalla sezione *IT/IS*.

3.4. Correlazione con lo stipendio

Nel line chart [6] vengono mostrati i coefficienti di Spearman tra performance (*PerfScoreID*) e stipendio (*Salary*), tra stipendio e soddisfazione dei dipendenti (*EmpSatisfaction*), e tra performance e soddisfazione, relativi alle posizioni aziendali in cui si registrano variazioni di performance. Come già accennato, poiché il dato dello stipendio registra variazioni significative a seconda della professione svolta all'interno dell'azienda (*Position*), le possibili correlazioni con la performance sono state valutate a parità di posizione.

L'istogramma [7] mostra invece lo stipendio medio per ciascuna delle posizioni considerate nel grafico [6]. L'affidabilità di ciascun valore medio è stata testata rispetto alla singola distribuzione di cui quel valore medio è rappresentativo.

Osserviamo che i valori significativi degli indici di correlazione nel grafico [6] sono spesso associati alle professioni più remunerative. Il caso più evidente è quello del *Data Analyst*, in cui tutte e tre le correlazioni hanno valore positivo. Si può ipotizzare che lo stipendio incida in tal caso sulla soddisfazione del dipendente, la quale a sua volta risulterebbe in un incremento della sua prestazione lavorativa. Un andamento simile, sebbene meno marcato, può essere osservato anche nelle posizioni di *Software Engineer* e di *Area Sales Manager*.

4. Parità di genere

L'obiettivo di questa sezione è approfondire l'aspetto della parità di genere all'interno dell'azienda. In base ai dati a disposizione, si è scelto di valutare il gender balance, l'eventuale presenza di un divario salariale e la relazione tra stato civile e assenze.

Il grafico a torta [8] mostra il gender balance all'interno dell'azienda, ovvero la percentuale di occupazione femminile e maschile in rapporto al totale. Osservando il grafico, la distribuzione globale degli impiegati rispetto al genere risulta abbastanza omogenea, con una lieve prevalenza del personale femminile su quello maschile.

Nel grafico [9] (grafico a barre di tipo *stacked*) è stato invece rappresentato il gender balance nelle diverse sezioni aziendali; anche qui abbiamo riscontrato una discreta omogeneità, con l'evidente eccezione dell'*Executive Office*, composto interamente da donne (va tuttavia notato che l'*Executive Office* è composto da un unico dipendente, cioè il CEO). Gli uffici amministrativi (*Admin Offices*) e la sezione *IT/IS* presentano, rispettivamente, una lieve prevalenza di personale femminile e maschile.

4.1. Valutazione del divario salariale

In questa sezione si sono valutate le differenze di stipendio tra donne e uomini a parità di posizione – sono state selezionate, dunque, solo le posizioni che registrano una presenza di personale sia maschile che femminile. Poiché per il calcolo della differenza sono stati considerati gli stipendi medi divisi per posizione e genere, abbiamo preliminarmente verificato che la media di ciascuna distribuzione fosse affidabile.

Nel line chart [10] si analizza lo scarto di stipendio tra uomini e donne a parità di posizione. Il valore 0 nel grafico rappresenta l'assenza di variazioni nello stipendio. In presenza di uno scarto positivo, gli uomini guadagnano in media più delle colleghe donne, e viceversa. L'andamento del grafico presenta delle variazioni poco marcate sia in positivo che in negativo. L'unica eccezione è rappresentata dalla posizione di *Network Engineer*, che presenta una differenza di stipendio positiva molto più marcata: in questo caso, sono gli uomini a guadagnare più delle impiegate donne.

Abbiamo poi deciso di calcolare e rappresentare mediante grafico a barre [11] la percentuale rappresentata dal divario salariale tra uomini e donne rispetto allo stipendio medio per ogni posizione. In tal modo, si è potuto rendere conto di quanto fosse significativa la differenza di stipendio. Osservando il grafico [11], le nostre intuizioni risultano confermate: la posizione di *Network Engineer* è l'unica che presenta uno scarto stipendiale significativo tra impiegati e impiegate.

4.2. Stato civile e assenze da lavoro

I boxplot in questa sottosezione mettono in relazione lo stato civile degli impiegati (espresso dalla variabile binaria *MarriedID*) e le assenze da lavoro. Se nel grafico [12] le donne sposate registrano un maggior numero di assenze rispetto alle colleghe single, la tendenza è invertita nel grafico [14]: sono gli impiegati sposati ad assentarsi di meno dal lavoro.

5. Diversità etnica

In questa sezione abbiamo indagato l'eterogeneità etnica degli impiegati nell'azienda e le eventuali relazioni dell'etnia con lo stipendio.

5.1. Distribuzione del personale rispetto all'identità etnica

Nel grafico a torta [16] si può osservare la distribuzione dei dipendenti rispetto alla loro identità etnica (autodichiarata). I dipendenti bianchi (*White*) sono maggioritari, ma anche il gruppo *Black or African American* ha una presenza consistente, seguito dai dipendenti asiatici (*Asian*) e da quelli ispanici (*Hispanic*). Minoritari sono i gruppi composti da dipendenti che si identificano come multirazziali (*Two or more races*).

Nel grafico a barre [17], invece, viene mostrata la composizione dei diversi dipartimenti aziendali rispetto all'etnia dei dipendenti, che non è uniforme. L'Executive Office appare la sezione meno diversificata, ma questo è dovuto al fatto che impiega un solo dipendente. Le sezioni composte da più gruppi etnici sono quelle di *Production* e *Sales*. *Production*, nonostante sia una delle sezioni più rappresentative (contiene infatti tutti i gruppi etnici), presenta anche la quota più alta di dipendenti bianchi.

Come già anticipato (par. 1.2), i dipendenti che hanno dichiarato di appartenere alla cultura ispanica sono stati raccolti nel gruppo *Hispanic*: questo perché si è scelto di dare prevalente importanza all'elemento culturale. Per completezza, nel grafico a torta [18] abbiamo illustrato la composizione etnica dei dipendenti che hanno dichiarato di appartenere alla cultura ispanica.

5.2. Stipendio e identità etnica

In questa sottosezione è stato analizzato il rapporto tra l'etnia dei dipendenti e il loro stipendio; abbiamo utilizzato il grafico a barre [19] per mostrare la distribuzione dello stipendio medio annuo rispetto all'etnia nelle varie sezioni aziendali. Dalla rappresentazione si può osservare un buon livello di omogeneità, alterato solo dal dipartimento *Executive* – che però, come già detto, si compone solamente del CEO.

Abbiamo inoltre verificato l'affidabilità dei valori medi per ciascuna delle distribuzioni, rilevando che per alcune distribuzioni la media affidabile (cioè senza outliers) sarebbe leggermente inferiore alla media mostrata nel grafico [19]. Ciononostante, in nessuna sezione aziendale si riscontrano differenze significative nello stipendio dei gruppi etnici. In aggiunta, è stato calcolato lo stipendio medio dei dipendenti raggruppati per etnia – previa rimozione degli outliers nei casi in cui la media della distribuzione dello stipendio non è risultata affidabile (cfr. Jupyter Notebook).

6. Terminazione del contratto

Da un'analisi della distribuzione dei dipendenti attivi e non più attivi [20] è emerso che il 33.4% dei lavoratori non fa più parte dell'azienda; per questo motivo abbiamo dedicato l'ultima parte della nostra analisi al tema del fine contratto.

6.1. Analisi delle cause di licenziamento

Gli ex-dipendenti sono stati suddivisi in dipendenti licenziati con motivazione e in dipendenti che hanno presentato dimissioni volontarie, come illustrato nel grafico a torta [21].

Nei grafici [22] e [23] sono mostrate le distribuzioni delle variabili *MotiviLicenziati* e *MotiviDimessi*, che descrivono rispettivamente le cause di licenziamento e quelle di dimissioni (par. 1.2).

Nei boxplot [24] e [25], invece, si possono osservare i livelli di engagement, soddisfazione e performance sia per i dipendenti dimessi volontariamente che per quelli licenziati. Confrontando i grafici della colonna [24] si può osservare come i dipendenti licenziati per motivi di performance abbiano anche un livello di soddisfazione inferiore alla media, cosa che ci ha portato ad approfondire la correlazione tra performance e soddisfazione dei dipendenti con performance inferiore alla media.

Dopo aver verificato la normalità delle distribuzioni dei dipendenti licenziati con motivazione e i dipendenti con performance inferiore alla media, abbiamo rappresentato tramite gli *scatter plot* [26] e [27] la correlazione tra *EmpSatisfaction* e *PerfScoreID* per entrambe le categorie (utilizzando il coefficiente di Spearman). Il grafico [26] mostra come per i dipendenti licenziati con motivazione non sussista una correlazione significativa tra grado di soddisfazione e livello di performance, mentre il grafico [27] mostra una debole correlazione positiva nei dipendenti con performance inferiore alla media.

6.2. Relazione tra durata dei contratti e stipendio degli ex-dipendenti

I boxplot [28] e [29] restituiscono, rispettivamente, lo stipendio medio in ogni sezione aziendale e la durata media del contratto per sezione aziendale. Abbiamo riscontrato che le sezioni *IT/IS* e *Production* presentano dati opposti: in *IT/IS* osserviamo stipendi alti, ma una durata media breve dei contratti; in *Production* osserviamo stipendi medi bassi, ma contratti più lunghi.

Negli scatter plot [28] e [29] abbiamo approfondito l'analisi della correlazione tra durata del contratto e stipendio medio nei dipartimenti *IT/IS* e *Production* (utilizzando il coefficiente di Spearman poiché le distribuzioni non sono risultate normali). Se la sezione *IT/IS* registra una correlazione positiva significativa, per quella di *Production* la correlazione risulta debolmente negativa.