Primo progetto R-Markdown

Cesareo Giacomo

Contents

Introduzione

Il mondo delle risorse umane è soggetto, come svariati altri settori, ad evoluzioni costanti. Questo cambiamento è dovuto principalmente, all'evoluzione tecnologica e alla conseguente esplosione dei Big Data. Questi due fattori hanno portato alla nascita degli HR Analytics, cioè, l'applicazione di sofisticati processi di Business Analytics, tecniche volte ad ottimizzare e prevedere gli andamenti di business e i risultati, alle risorse umane.

Obiettivo dell'HR Analytics è quello di fornire e analizzare le statistiche necessarie per gestire con efficienza le risorse umane all'interno di una organizzazione aziendale. Per questo motivo, l'interesse di avere un Data Scientist all'interno degli uffici risorse umane delle aziende sta crescendo di giorno in giorno. Questo per cercare, attraverso l'analisi dei dati, di rispondere alle domande: Chi sarà un mio dipendente fedele? Chi cambierà lavoro? Quali attributi personali spingono verso la prima o la seconda direzione?
Il Dataset utilizzato per questa relazione presenta informazioni relative a 1470 dipendenti che vengono descritti in base a 35 variabili.

Per rispondere alla precedente domanda, sarà tenuta maggiormente in considerazione nella prima parte di questa analisi, la variabile Attrition: un problema tipico di ogni azienda aziende. Esso comporta costi significativi per un'azienda soprattutto nella fase di assunzione e formazione di nuovo personale. Per questo motivo vi è un grande interesse, da parte dell'azienda, ridurre al minimo tale fenomeno. Individuare le cause dell'attrition permetterebbe alle Risorse Umane di intervenire in tempo al fine di evitare l'abbandono da parte dei propri dipendenti.

L'obbiettivo di questa analisi è quello andare a trattare la maggior parte degli argomenti svolti a lezione, per questo l'analisi si dividerà in tre parti:
1. La prima parte sarà incentrata su statistiche descrittive, al fine di comprendere com'è strutturata l'azienda, con l'obbiettivo di capire se vi sono e quali sono le variabili che influenzano l'abbandono dei dipendenti. Verrà inoltre trattata la variabile MonthlyIncome, vale a dire il reddito dei dipendenti, con lo scopo di capire come questa variabile cambi in relazione alle altre variabili per stabilire se possano esserci delle connessioni con le altre variabili.
2. La seconda parte tratta l'ANOVA, quindi la parte della statistica inferenziale al fine di valutare se vi sono differenze sostanziali tra le medie del reddito dei dipendenti suddivisi in gruppi considerando le variabili **JobRole**, che sta ad indicare appunto le varie professioni interne, e **MaritalStatus**, che rappresenta i diversi stati civili dei dipendenti.
3. La terza parte, ed ultima parte, è dedicata alla regressione, con il fine di prevedere diversi valori, sempre per quanto riguarda la variabile **MonthlyIncome**, di salario dei dipendenti in base alle variabili **Age** e **TotalWorkingYears**.

Il Dataset presenta sia dati di tipo qualitativo che quantitativo. Nello specifico, in questo elaborato verranno utilizzate le seguenti variabili:
• Qualitativi nominali: **Attition** (binaria), **Gender** (binaria), **MaritalStatus**, **JobRole**, **OverTime** (binaria).
• Qualitativi ordinali: **JobSatisfaction**.
• Quantitativi discreti: **Age**, **TotalWorkingYears**.
• Quantitativi continui: **MonthlyIncome**.

1. Statistiche Descrittive

In questo primo capitolo viene effettuata un'analisi descrittiva dei dati, in modo tale da far emergere quali sono gli aspetti interni all'impresa sui quali occorre soffermarci, rispetto alle 35 le variabili che vengono fornite dal Dataset. Si vuole quindi rendere più chiaro il contesto e fornire delle linee guida per analisi successive.

1.1 Variabili Attrition e Gender

Per prima cosa si va ad investigare come viene suddivisa la popolazione di riferimento in base alla variabile Attrition, che sta ad indicare se il dipendete ha cessato o meno il rapporto con l'impresa in questione. Nella seguente tabella vengono presentate le frequenze assolute e relative della variabile Attrition per vedere come questa si distribuisce nell'organico dell'impresa: l'84% circa dei dipendenti non ha cessato il rapporto mentre il 16% degli stessi ha lasciato l'impresa. Si può notare come questo non è assolutamente un valore trascurabile, dal momento che per ogni dipendente viene eseguito un investimento monetario, soprattutto per quanto riguarda la ricerca del personale e della sua formazione, quindi verranno presentate in seguito quali potrebbero essere state le cause della cessazione del rapporto lavorativo.

Attrition	Count	Frequenze Relative
No	1,233	0.8388
Yes	237	0.1612

Viene presentata, come si può vedere in *figure 1*, la visualizzazione grafica delle frequenze assolute della tabella precedente attraverso l'utilizzo di un countplot.

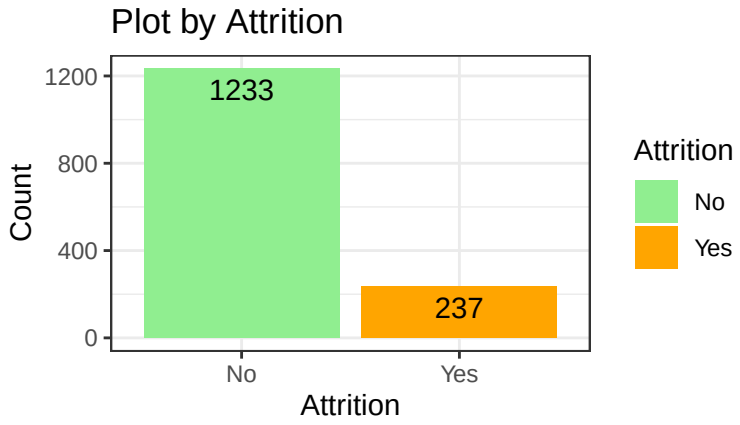


Figure 1

Ora che è stato esposto quanti dipendenti hanno abbandonano l'azienda si vuole indagare su quali siano state le motivazioni che hanno portato questi ultimi a prendere questa decisione.
Come prima variabile viene presa in considerazione **Gender** sulla quale sono state eseguite le stesse statistiche descrittive utilizzate per la variabile **Attrition**.
Per quanto riguarda le frequenze assolute e le frequenze relative possono esssere osservate nella tabella sottostante.

Gender	Count	Frequenze Relative
Female	588	0.4
Male	882	0.6

Si evince una sostanziale maggioranza di dipendenti di sesso maschile (60%) rispetto a quello femminile (40%).
Viene ora presentata una rappresentazione grafica (*figure 2*) della variabile **Gender** sempre tramite l'utilizzo di un countplot.

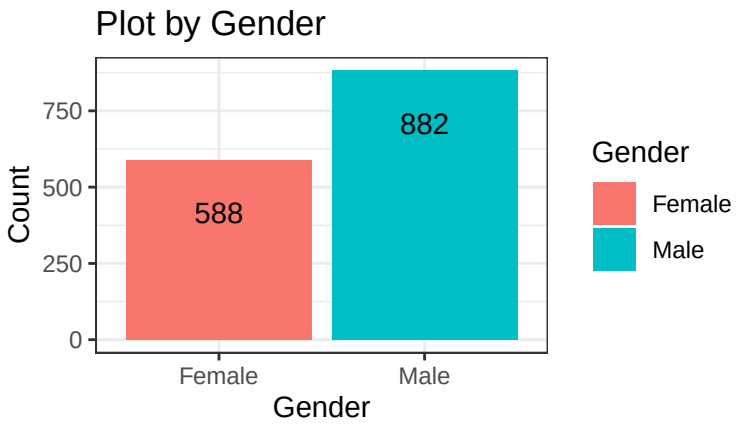


Figure 2

1.1.1 Analisi bivariata

Vengono ora studiati, congiuntamente, questi due caratteri presentati attraverso l'utilizzo della tabella a doppia entrata seguente.

	No	Yes
Female	501	87
Male	732	150

Questa tabella bivariata fornisce solo valori assoluti, difficilmente interpretabili: mostra solo quanti uomini e quante donne hanno cessato o meno il rapporto con l'azienda. Sarebbe più utile infatti osservare la seguente tabella bivariata che rappresenta le frequenze relative percentuali della variabile **Attrition** relazionata alla variabile **Gender**.

	No	Yes
Female	0.852	0.148
Male	0.8299	0.1701

Si evince infatti come, in termini percentuali, sono gli uomini ad avere una più alta frequenza di abbandono dell'impresa, comunque una differenza molto bassa (2.21%). Considerando invece la stessa tabella bivariata delle frequenze assolute, rapportata al numero totale di osservazioni, essa non avrebbe dato informazioni rilevanti su quale genere sia in realtà quello con una probabilità maggiore di Attrition.
Di seguito (*figure 3*) viene presentata la visualizzazione delle frequenze assolute della tabella osservata precedentemente, così da ottenere una sintesi dei risultati ottenuti.

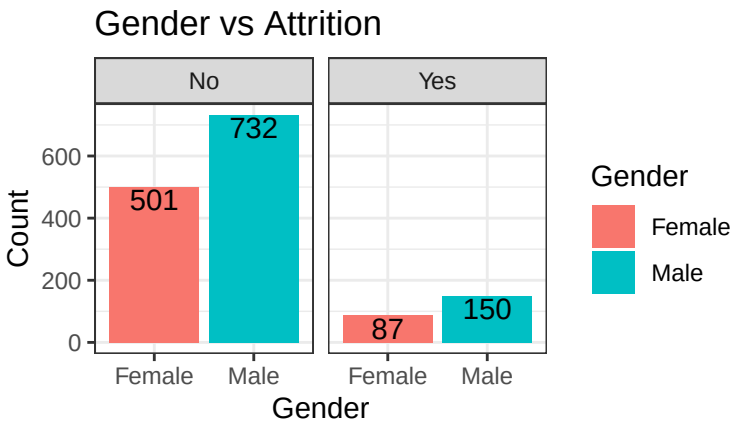


Figure 3

1.1.2 Connessione tra Attrition e Gender

Dopo aver analizzato singolarmente e congiuntamente le due variabili, si propone ora, grazie all'utilizzo del test chi-quadro, un'analisi volta a valutare se tra queste due variabili è presente un'eventuale relazione. Se non esiste alcuna relazione si dirà che le due variabili sono statisticamente indipendenti.
Per valutare il grado di conessione tra due caratteri qualitativi viene utilizzato l'indice *Chi-quadro di Pearson*, ma, dal momento che questo valore non è interpretabile, sarà necessario ricorrere alla sua normalizzazione: l'indice normalizzato varia da zero, assenza di connessione (indipendenza statistica) ad 1, massima connessione (ad ogni modalità di Gender corrisponde una ed una sola modalità di Attrition). Come si può osservare dalla tabella sottostante, nel nostro caso il valore è prossimo allo zero e quindi concludiamo che le variabili sono indipendenti.

ChiQuadro	ChiQuadroNorm
1.117	0.0007598

Table 6: Pearson's Chi-squared test with Yates' continuity correction: `data$Gender` and `data$Attrition`

Test statistic	df	P value
1.117	1	0.2906

Anche osservando il p-value (0.2906) del Test Chi-Quadro non si sarebbe rifiutata l'ipotesi nulla di indipendenza statistica.

Il grafico seguente (*Figure 4*) misura la magnitudine con la quale i residui impattano sul valore del chi-Quadro. I residui standardizzati, ovvero le contingenze rapportate alla radice quadrata delle frequenze teoriche, sono importanti per interpretare l'associazione tra le righe le colonne della tabella bivariata: quelli in rosso sono positivi, rappresentano un'attrazione tra le variabili di riga e quelle di colonna, ad esempio nel nostro caso troviamo un alta associazione positiva tra *Yes* e *Male*, mentre quelli di colore blu rappresentano residui negativi, nel nostro caso tra *Female* e *Yes*, quindi una forte repulsione.

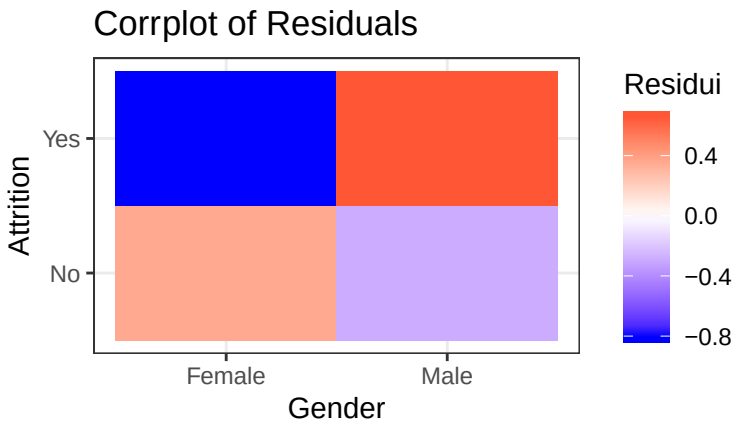


Figure 4

Si può solo afermare indipendenza tra queste due variabili, tuttavia il carattere qualitativo **Gender** verrà utilizzato in analisi successive e messo in relazioni con altre variabili, dato che potrebbe mettere in risalto aspetti importanti, come ad esempio la differenza di reddito (**MonthlyIncome**) tra i due sessi.

1.2 Variabile MonthlyIncome

Si andrà ora ad analizzare la variabile MonthlyIncome, sempre in relazione alla variabile Attrition, per valutare se la causa dell'allontanamento dei dipendenti dall'azienda sia da attribuire ad una retribuzione non sufficiente.
Per prima cosa, dato che la variabile MonthlyIncome è di tipo quantitativo continuo, può essere utile visualizzare queste due variabili tramite l'utilizzo sia di un boxplot sia di un violin-plot (*Figure 5*) per analizzare come si distribuisce in relazione alla variabile Attrition.

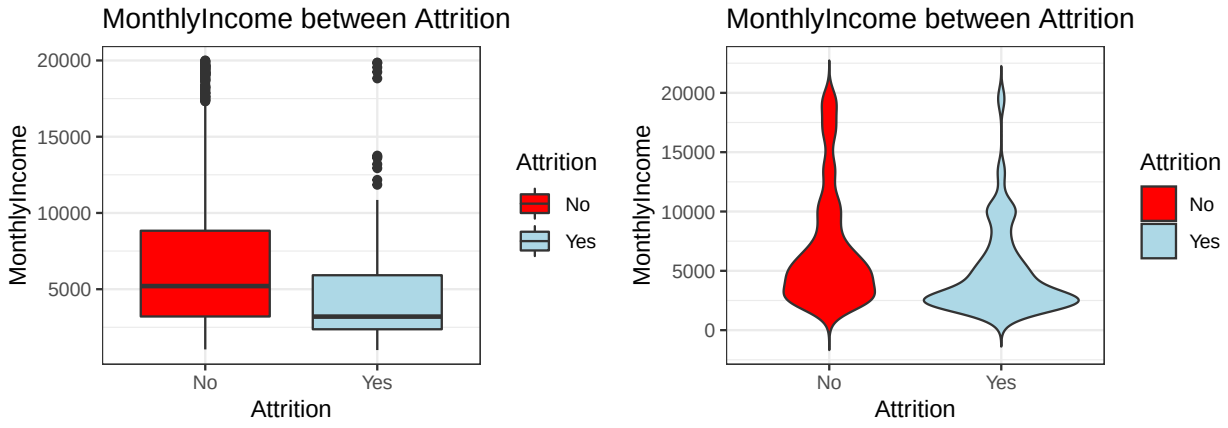


Figure 5

Dal boxplot si evince una sostanziale differenza tra le due distribuzioni: la mediana risulta inferiore per quanto riguarda i dipendenti che hanno abbandonato l'impresa rispetto a coloro che sono rimasti fedeli a quest'ultima, il range interquartile risulta più schiacciato comportando così una deviazione standard minore. Osservando il violin-plot si nota, per quanto concerne coloro che non hanno abbandonato la compagnia, una distribuzione di probabilità più omogenea rispetto all'altro gruppo. Si nota comunque una densità più elevata per valori di **MonthlyIncome** compresi tra 2.500 e 5.000 dollari. Totalmente diversa è invece la distribuzione dell'altro gruppo: si nota un'altissima concentrazione intorno al valore di 2.500 dollari. Da entrambi i grafici si osservano outliers nella parte superiore della distribuzione, in particolare per coloro che hanno cessato i rapporti con la compagnia: si osservano dei punti intorno ai 20.000 che avranno avuto sicuramente altre motivazioni per aver lasciato l'impresa.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Yes Attrition	1,051	3,211	5,204	6,833	8,834	19,999
No Attrition	1,009	2,373	3,202	4,787	5,916	19,859

Dalla tabella precedente si ricavano dati più precisi ed esaustivi: le differenza tra le medie e le mediane dei due gruppi sono molto alte, rispettivamente di 2.002 e 2.046 dollari, ma la differenza maggiore si ha osservando il terzo quartile dei due gruppi con uno scarto di 2.918 dollari. Quindi il 75% del primo gruppo percepirà un salario uguale o minore a 5.916 dollari mentre il 75% del secondo gruppo uguale o minore a 8.834.
Si può dunque affermare che il salario sarà stata una delle variabili maggiormente considerate da coloro che hanno lasciato l'impresa.

1.3 Variabili MonthlyIncome e Gender

Di seguito viene proposta un'analisi della variabile **MonthlyIncome** relazionata al carattere **Gender**.
Come è purtroppo noto, nella maggioranza dei casi, le donne percepiscono, mediamente, un salario minore rispetto agli uomini. Viene riportata una tabella di sintesi delle statistiche descrittive tra le due variabili.

Gender	Salary Mean	Salary Median	Standard Deviation
Female	6,687	5,082	4,696
Male	6,381	4,838	4,715

Si può notare una disparità inversa per quanto riguarda la distribuzione dei salari rispetto a quanto affermato precedentemente: nella compagnia le donne percepiscono, mediamente, uno stipendio lievemente maggiore rispetto ai colleghi di sesso maschile (lo si può osservare anche confrontando la mediana tra i due gruppi dato che ques'ultima è meno sensibile ai valori estremi). La deviazione standard è molto alta data la grande variazione di valori presenti nei dati.

1.3.1 Discretizzazione variabile MonthlyIncome

Per praticità si opera ad una discretizzazione della variabile **MonthlyIncome**, dato che è espresso su scala continua, esprimendo i livelli assunti dalla variabile in categorie intervallari, ognuna delle quali di ampiezza uguale a 1.000 dollari. Di seguito viene presentata la tabella bivariata con annesse le frequenze relative dei due gruppi.

	Female	Male	Relative Female	Relative Male
[1 - 2]	10	23	0.01701	0.02608
[2 - 3]	136	226	0.2313	0.2562
[3 - 4]	50	97	0.08503	0.11
[4 - 5]	93	114	0.1582	0.1293
[5 - 6]	68	97	0.1156	0.11
[6 - 7]	50	71	0.08503	0.0805
[7 - 8]	23	31	0.03912	0.03515
[8 - 9]	19	29	0.03231	0.03288
[9 - 10]	16	36	0.02721	0.04082
[10 - 11]	25	41	0.04252	0.04649
[11 - 12]	10	10	0.01701	0.01134
[12 - 13]	5	5	0.008503	0.005669

	Female	Male	Relative Female	Relative Male
[13 - 14]	26	17	0.04422	0.01927
[14 - 15]	2	7	0.003401	0.007937
[15 - 16]	4	3	0.006803	0.003401
[16 - 17]	17	12	0.02891	0.01361
[17 - 18]	10	18	0.01701	0.02041
[18 - 19]	10	9	0.01701	0.0102
[19 - 20]	14	36	0.02381	0.04082

Da questa tabella si osserva come le classi più popolate per entrambi i sessi sono quelle relative agli intevalli compresi tra i 2.000 e i 3.000 dollari e tra i 4.000 e i 5.000 dollari. Un'altra tabella utile per la visualizzazione dei quartili è quella relativa alle frequenze relative cumulate.

	Cumulate Female	Cumulate Male
[1 - 2]	0.01701	0.02608
[2 - 3]	0.2483	0.2823
[3 - 4]	0.3333	0.3923
[4 - 5]	0.4915	0.5215
[5 - 6]	0.6071	0.6315
[6 - 7]	0.6922	0.712
[7 - 8]	0.7313	0.7472
[8 - 9]	0.7636	0.78
[9 - 10]	0.7908	0.8209
[10 - 11]	0.8333	0.8673
[11 - 12]	0.8503	0.8787
[12 - 13]	0.8588	0.8844
[13 - 14]	0.9031	0.9036
[14 - 15]	0.9065	0.9116
[15 - 16]	0.9133	0.915
[16 - 17]	0.9422	0.9286
[17 - 18]	0.9592	0.949
[18 - 19]	0.9762	0.9592
[19 - 20]	1	1

Si osserva come per gli uomini il primo quartile risieda nell'intervallo [2-3], per le done invece lo stesso risiede, per uno scarto davvero basso, nell'intervallo [3-4]. Essendo comunque una differenza davvero bassa per le donne si potrebbe arrotondare per eccesso e considerare dunque il primo quartile nell'intervallo [2-3] anche per queste ultime. Stesso discorso per il secondo quartile: si può considerare appartenente alla classe [4-5] per entrambi i gruppi. Il terzo quartile appartiene alla classe [8-9] per entrambi i generi. Considerando appunto che lo stipendio può variare da un minimo di 1.000 ad un massimo di 20.000 si può affermare che questa differenza tra i salari medi di genere femminile rispetto a quelli di genere maschile è trascurabile.

Si propone ora grazie ai seguenti due grafici come si distribuiscono le due variabili.

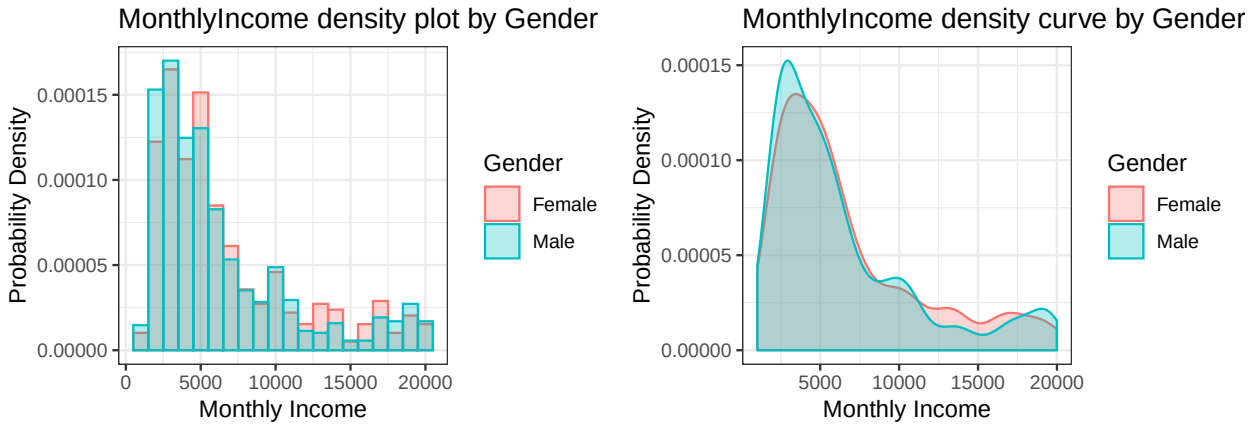


Figure 6

Nel grafico a sinistra sono stati scelti *bins* di ampiezza 1.000 dollari e si può notare la superiorità nella distribuzione di probabilità del salario delle donne. Nel grafico di destra che rappresenta le curve di densità di probabilità questa superiorità risulta molto più evidente dato che la curva relativa alle donne risulta superiore per la maggioranza della distribuzione (si era gia osservato dalle frequenze relative).

1.3.2 Connessione tra MonthlyIncome e Gender

Utilizzando i dati relativi alle tabelle bivariate precedenti, si può ora effettuare un test di chi-quadro per valutare il grado di connessione tra le variabili **Gender** e **MonthlyIncome**.

ChiQuadro	ChiQuadroNorm
28.25	0.01922

Il valore del chi quadro normalizzato è molto basso il che suggerisce che le due variabili sono tra loro indipendenti.

Table 12: Pearson's Chi-squared test: `tablebinned1`

Test statistic	df	P value
28.25	18	0.05834

Anche osservando il P-value si nota come questo sia di poco superiore al livello critico dello 0.05, si può quindi accettare l'ipotesi nulla di indipendenza tra le due variabili.

Dato che la variabile **MonthlyIncome** è di tipo quantitativo si potrebbe anche calcolare l'indipendenza in media tramite l'indice η^2. Ci si aspetta un valore di η^2 prossimo allo zero dato che l'indipendenza in distribuzione implica l'indipendenza in media.

Eta.Quadro
0.001015

Come appena accennato tra le due variabili non vi è indipendenza in media dato che il valore di η^2 è vicino a zero, infatti può assumere solo valori compresi tra 0 ed 1: vale zero se la varianza fra i gruppi è nulla cioè quando **MonthlyIncome** è indipendente in media da **Gender** (e la varianza nei gruppi coincide con la varianza marginale di **MonthlyIncome**), mentre vale 1 quando la varianza fra i gruppi coincide con la varianza marginale di **MonthlyIncome** cioè quando **MonthlyIncome** è perfettamente dipendente da **Gender** (e la varianza nei gruppi è nulla).

1.4 Connessione Attrition JobSatisfaction

Molto spesso alla base dell'Attrition c'è un malcotento del dipendente. Proprio per questo l'azienda dovrebbe monitorare la JobSatisfaction dei propri dipendenti, per capire come intervenire per tempo lì dove quest'ultimo manifesta la propria insoddisfazione. A tal fine, decidiamo di analizzare la variabile JobSatisfaction.

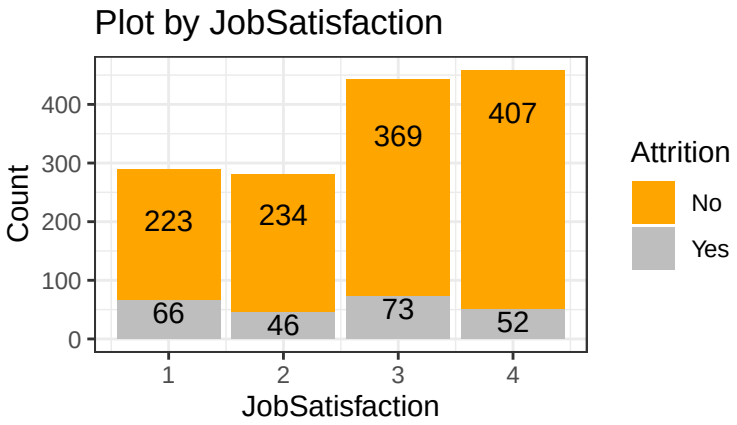


Figure 7

Vediamo che le persone che abbandonano l'azienda hanno un livello di soddisfazione molto basso. E' possibile supporre che il dipendente decide di andar via a causa dell'insoddisfazione. Questi insight e quelli del punto precedente (monthly income), possono spingerci ad approfondire se l'insoddisfazione dei dipendenti è legata soprattutto alla retribuzione. Se così fosse, possiamo confermare che i dipendenti nella maggior parte dei casi lasciano l'azienda perché insoddisfatti dal punto di vista retributivo.

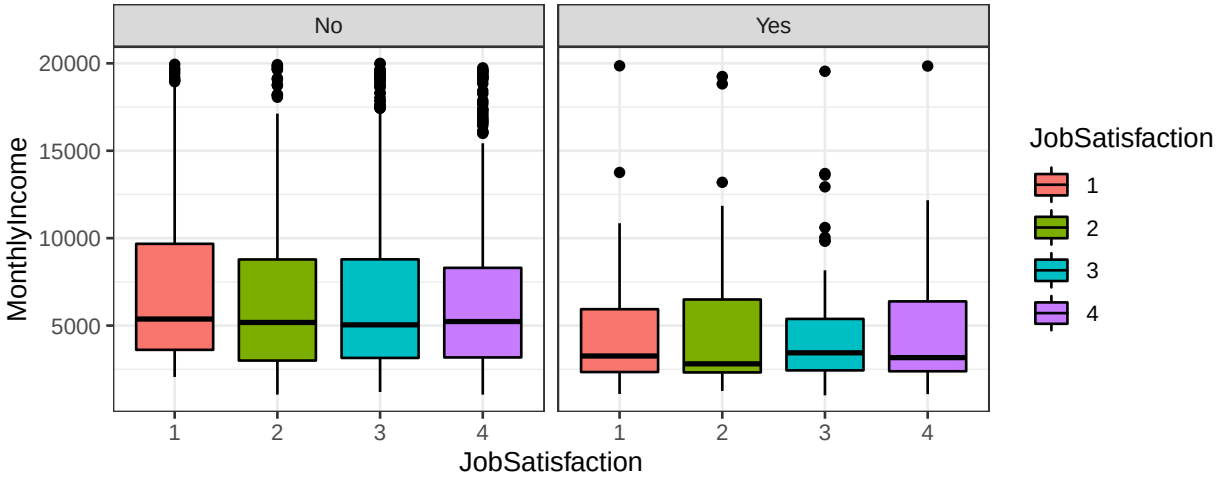


Figure 8

Il grafico sopra (*figure 8*) ci mostra la distribuzione dei dipendenti per JobSatisfaction/MonthlyIncome/Attrition. Appare evidente che, a parità di JobSatisfaction, le persone che abbandonano l'azienda hanno uno stipendio più basso rispetto a coloro che non hanno abbandonato l'azienda.

1.5 Variabili Attrition Overtime

Accade di frequente che nelle aziende i dipendenti decidono di cambiare lavoro a causa del troppo tempo che sono costretti a dedicare a quest'ultimo. In molti casi i dipendenti oberati sono costretti a rimanere sul luogo lavorativo oltre le ore stabilite da contratto per far fronte alla mole di lavoro. Si vuole analizzare se anche per l'azienda in questione c'è un legame tra abbandono e "straordinari". Per far questo, procediamo con l'analisi bivariata tra Attrition e Overtime.

	Not Overtime	Overtime
No	944	289
Yes	110	127

	Not Overtime	Overtime
No	0.8956	0.6947
Yes	0.1044	0.3053

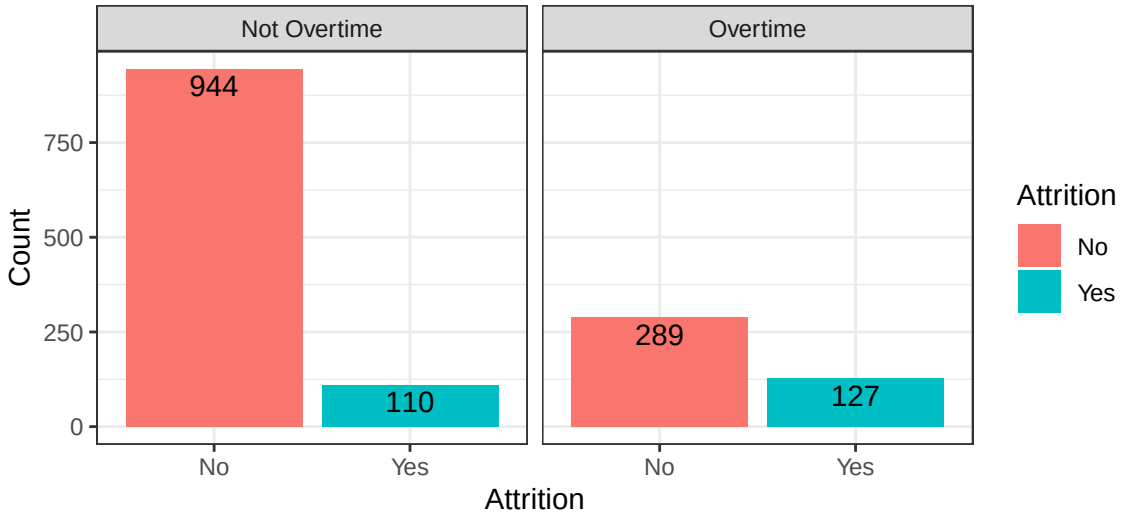


Figure 9

Dalle tabelle e dal grafico sopra (*figure 9*), si evince che c'è Overtime nel gruppo, ovvero, tra coloro che lavorano più delle ore previste da contratto, c'è un tasso di abbandono di circa 3 volte superiore di quello del gruppo Not Overtime. Si può dedurne quindi che le persone probabilmente abbandonano l'azienda anche per l'eccessivo carico di lavoro.

2. ANOVA

In questa sezione vengono proposte due analisi della varianza (ANOVA): nella prima verrà suddivisa la popolazione in base alla variabile **MaritalStatus**, che sta ad indicare lo stato civile dei dipendenti della compagnia, mentre nella seconda si suddividerà la popolazione in gruppi discriminati secondo la variabile **JobRole**, che indica appunto il tipo di lavoro svolto da questi ultimi.

L'ANOVA si basa sulla valutazione delle medie di queste gruppi, ottenuti tramite la suddivisione in base a queste due variabili categoriali, e si vuole verificare se queste sono uguali tra di loro. L'ipotesi nulla afferma che le medie di tutti i gruppi sono uguali tra di loro mentre l'ipotesi alternativa afferma invece, se accettata, che anche solo una di queste medie differisce dalle altre.
Viene utilizzata l'anova perchè si vuole valutare la media di più di due popolazioni, altrimenti si sarebbe utilizato un *t-test*, in modo da valutare come interagiscono quantitativamente, tra di loro, più di due gruppi.

2.1 ANOVA MonthlyIncome ~ MaritalStatus

Come è stato accennato precedentemente verrà utilizzata la variabile **MaritalStatus** in modo da suddividere la popolazione in tre gruppi: "Married", "Divorced" e "Single" in modo da poter verificare se le medie del reddito di questi tre gruppi sono uguali tra di loro.

Come prima cosa è utile mostrare le frequenze assolute e relative in base allo stato civile di ciascuno dei dipendenti.

Divorced	Married	Single
327	673	470

Divorced	Married	Single
0.2224	0.4578	0.3197

Si nota come la maggioranza dei dipendenti risulti sposata, quasi il 46%, i *Single* rappresentano circa il 32% del totale mentre i divorziati rappresentano la classe meno popolata rappresentata da circa il 22% del totale. Dato questo elevato numero di *Single* sarebbe interessante chiedersi se questi ultimi siano più giovani rispetto agli altri individui, se così fosse percepirebbero, mediamente, un salario minore rispetto agli altri colleghi dato che possiedono un'esperienza lavorativa minore. Per quanto riguarda invece sposati e i divorziati, questi avranno sicuramente qualche anno in più rispetto ai colleghi *Single* e ci si aspetta dunque di osservare un salario maggiore per questi ultimi.

Nella seguente tabella sono riportati i quartili, il minimo ed il massimo degli stipendi suddivisi in base allo stato civile.

	Min.	First.Qu.	Median	Mean	Third.Qu.	Max.
Divorced	1,129	3,015	5,131	6,786	9,418	19,973
Married	1,052	3,022	5,204	6,794	9,096	19,999
Single	1,009	2,722	4,536	5,889	7,328	19,926

Si può ora affermare quanto ipotizzato precedentemente: i *Single* guadagnano di meno sia osservando la media che la mediana. Ricordando che la media del salario dell'intera popolazione è di 6.503 dollari si evince come le medie di sposati e divorziati non sono cosi lontane dalla media totale e sono quasi uguali tra di loro. Si può anche affermare che il massimo per tutti e tre i gruppi sia grossomodo lo stesso, mentre per quanto riguarda il minimo si osserva solo come i divorziati abbiano uno stipenio minimo poco superiore ai loro colleghi sposati e *Single.* Guardando invece al terzo quartile si nota una sostanziale differenza per quanto riguarda tutte e tre le categorie: il 75% dei *Single* percepisce un salario uguale o minore a 7.328 dollari, più di 2000 dollari in meno rispetto ai colleghi divorziati. Come è stato detto precendetemente, potrebbe essere causato dalla maggiore inesperienza dei giovani rispetto ai loro colleghi più anziani. Si osserva anche una differenza di quasi 400 dollari tra divorziati e sposati. Questa discrepanza potrebbe essere dovuta al fatto che i dipendenti divorziati potrebbero avere più tempo da dedicare al lavoro rispetto ai colleghi sposati che hanno comunque del tempo da

dedicare alla famglia.

Per comodità viene proposta una rappresentazione grafica dei tre gruppi grazie ad un Boxplot che permette anche di valutare la presenza di outliers all'interno della popolazione.

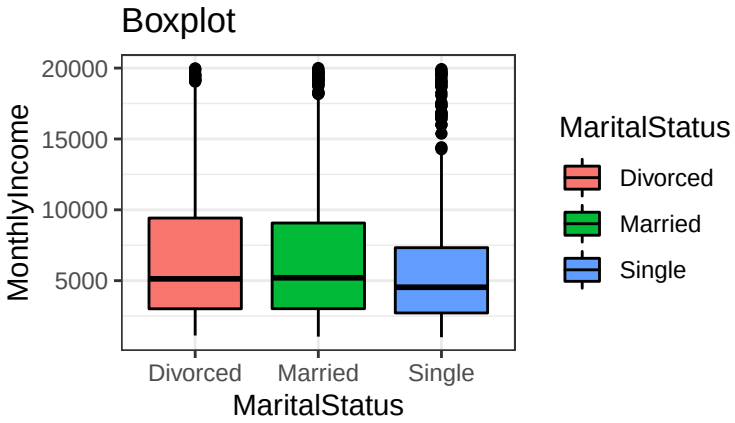


Figure 10

Dall'analisi grafica del Boxplot si evince quello che è stato gia detto in precedenza: tra i gruppi *Divorced* e *Married* minimo e mediana risultano quasi identici mentre il range interquartile è leggermente minore per quanto riguarda i dipendenti sposati. Interessnte è invece la distribuzione dei *Single*: presentano una differenza interquartile minore e, come visto nella tabella precendente, minore mediana. Dai valori minimi e massimi del boxplot notiamo come i *Single* hanno anche una deviazione standard minore rispetto gli sposati e i divorziati.
Notiamo anche come il salario minimo sia praticamente uguale per tutte e tre le classi e non ci siano outliers nella parte inferiore della distribuzione. Situazione diversa per quanto riguarda la parte alta della distribuzione dato che tutte e tre le classi presentano numerosi outliers.

Si procede ora con il calcolo degli outliers per classe dato che dal grafico non si riesce a comprendere appieno la numerosità di questi punti estremi.

Single	Married	Divorced
33	34	15

Per quanto riguarda i dipendenti *Single* si nota come outliers rappresentino circa il 10% della distribuzione del gruppo e, come si vede dal boxplot, alcuni di questi punti sono di molto fuori dal limite consentito. Queste osservazionii infatti potrebbero essere molto influenti sulla distribuzione totale.
Ci si aspetta, come ipotizzato precedentemente, che i *Single* siano i dipendenti più giovani dell'impresa, quindi, vengono proposti i boxplot e i violin-plot relativi alla variabil **Age** in relazione con lo stato civile, e una tabella rappresentante statistiche descrittive di base.

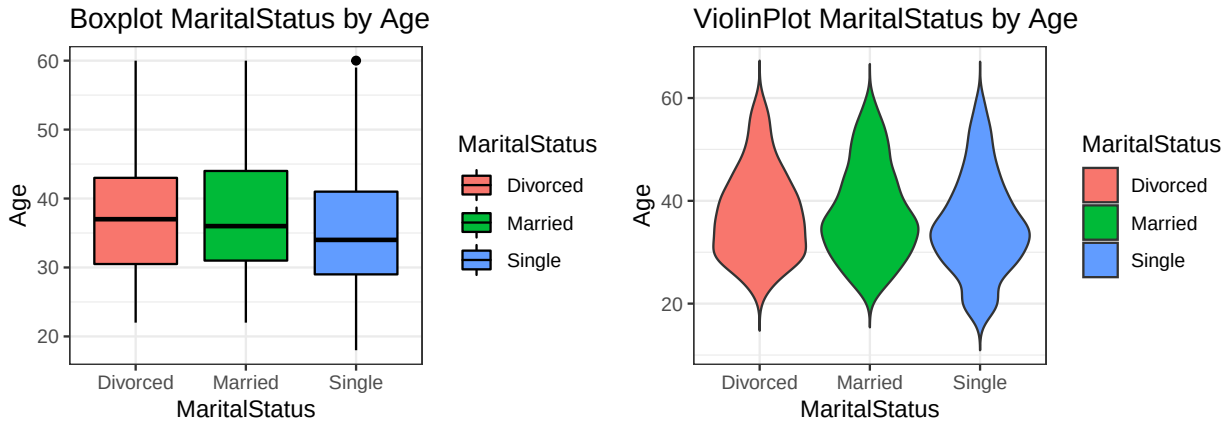


Figure 11

Marital.Status	Mean	Median	Standard.Deviation
Divorced	37.49	37	8.574
Married	37.76	36	9.003
Single	35.34	34	9.508

Osservando il boxplot si può ora confermare quanto detto in precedenza: i lavoratori *Single* sono in media più giovani dei dipendenti divorziati e Sposati. Questi presentano anche una deviazione standard maggiore, lo si evince anche dal grafico dato che presentano una coda inferiore più basse rispetto alle altre due classi. La situazione risulta ancora più chiara guardando il volin-plot: per quanto riguarda il gruppo relativo ai *single* troviamo una densità di probabilità maggiore nella parte inferiore della distribuzione rispetto agli altri due gruppi.

Tornando all'analisi dello stipendio viene proposta la visualizzazione grafica della densità di probabilità delle tre classi all'interno di un unico grafico.

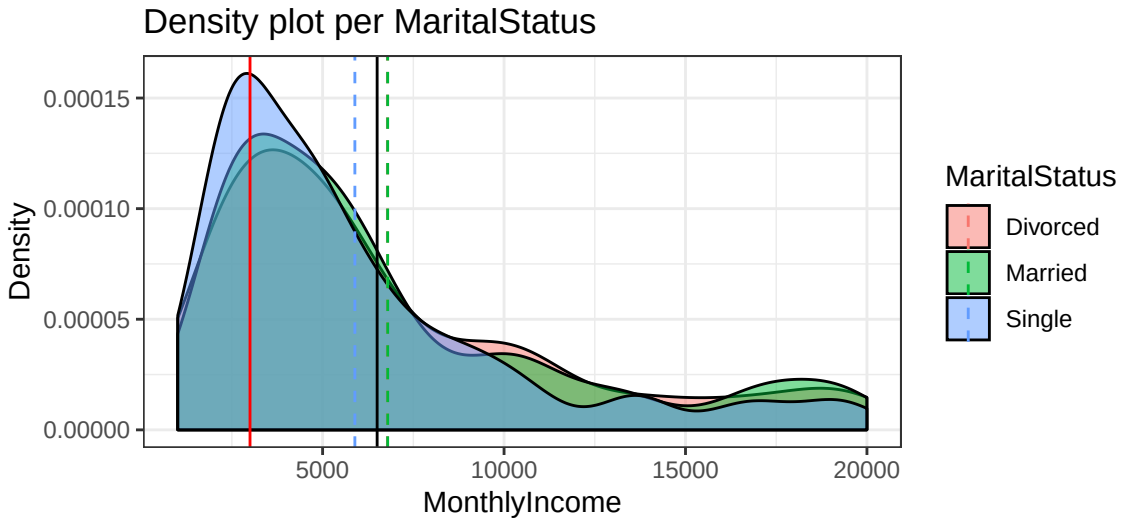


Figure 12

Si consideri ora solamente l'intervallo compreso tra 5.000 e 7.000 dollari dato che in questo primo grafico le medie dei dipendenti sposati e divorziati sono sovrapposte.

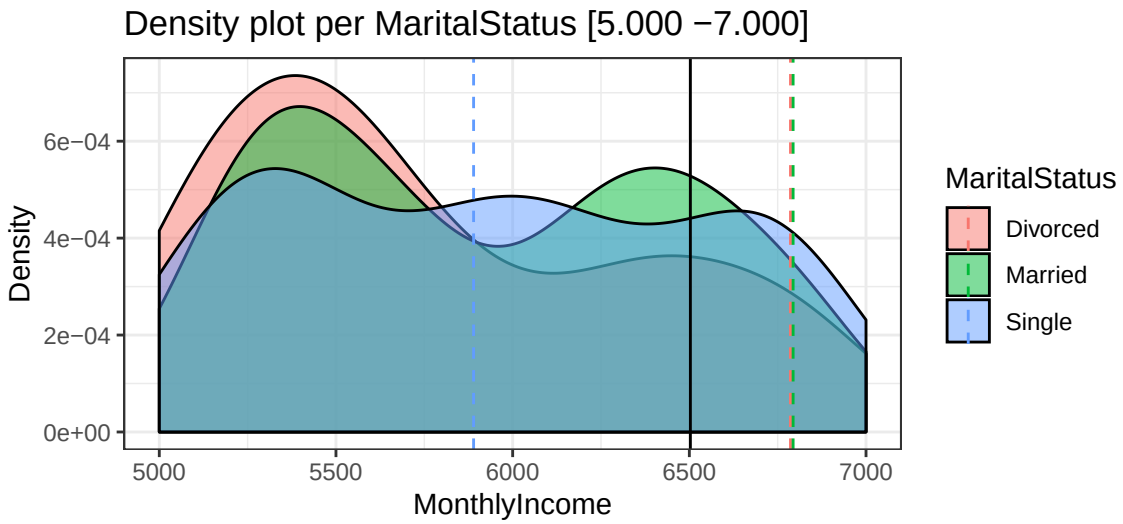


Figure 13

Dalla curva di probabilita della variabile **MonthlyIncome** si evince come la distribuzione presenti un'assimmetria positiva (Media < Mediana), si nota infatti un picco per quanto riguarda un valore prossimo a 3.000 dollari (linea rossa), soprattutto per quanto riguarda i dipendenti *Single.* La distribuzione presenta infatti una varianza molto elevata: una varianza nei gruppi molto ampia dato che il salario varia da circa 1.000 a 20.000 dollari, per quanto riguarda invece la varianza tra i gruppi si osserva come questa sia bassa per quanto riguarda il gruppo di *Single* ed ancora piu bassa per gli altri due gruppi dato che le due medie sono molto vicine alla media totale rappresentata dalla linea nera. Si era gia osservato in precedenza nella tabella suddivisa i classi di stipendio che più di un quarto della popolazione percepisce un reddito compreso da 0 a 3.000 dollari, infatti con una varianza così elevata questo è il classico caso dove media e anche mediana (anche se piu robusta) vengono trascinati da valori estremi.

La media dei *Single* sembra un po piu bassa e vogliamo valutare se è dovuto ad un fattora casuale. Dividiamo in gruppi in base allo stato matrimoniale, qundi avremo tre gruppi ed eseguiremo l'anova in relazione allo stipendio.
Si nota come le medie tra *Divorced* e *Married* si sovrappongono, la piu lontana dalla media totale è *Single.*

Vengono presentati ora gli output relativi all'Anova.

Table 21: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MaritalStatus	2	2.6e+08	1.3e+08	5.905	0.002792
Residuals	1,467	3.23e+10	22,017,829	NA	NA

il valore F-value, ovvero il valore critico, è il rapporto tra il *mean squared error* delle variabile esplicative (nel nostro caso solo **Maritalstatus**) e il *mean squared error* dei residui del modello. Il p-value indica invece la probabilità di ottenere un valore uguale o più estremo rispetto al valore F-value osservato. Considerando un livello di significatività uguale a 0.05 rifiutiamo l'ipotesi nulla di uguaglianza nelle medie dei nostri gruppi.
Affermiamo quindi che con i nostri dati non vi è una forte evidenza che le medie dei gruppi relativi alla variabile Marital-Status siano tra di loro uguali.

2.2 ANOVA MonthlyIncome ~ JobRole

Si vuole ora confrontare come tenda a variare il reddito in relazione alla variabile **JobRole**.

Per prima cosa vegono proposte statistiche descrittive preliminari per valutare come si distribuiscono la classe **JobRole** all'interno della popolazione di riferimento.

JobRole	Count	Frequenze Relative
Healthcare Representative	131	0.08912
Human Resources	52	0.03537
Laboratory Technician	259	0.1762
Manager	102	0.06939
Manufacturing Director	145	0.09864
Research Director	80	0.05442
Research Scientist	292	0.1986
Sales Executive	326	0.2218
Sales Representative	83	0.05646

Si osserva come il maggior numero di dipendenti occupino la posizione di Sales Executive seguiti da Research Scientist e Laboratory Technician, infatti queste tre professioni occupano circa il 60% dei dipendenti totali.

Vengono ora proposte le medie suddivise per categoria e si nota che è presente una varianza molto alta dato che in base al tipo di lavoro i salari cambiano di molti da un minimo di circa 2.500 a un massimo di 17.000 dollari.

JobRole	Salary Mean
Healthcare Representative	7,529
Human Resources	4,236
Laboratory Technician	3,237
Manager	17,182
Manufacturing Director	7,295
Research Director	16,034
Research Scientist	3,240
Sales Executive	6,924
Sales Representative	2,626

Incrociando i dati delle due tabelle precedenti si osserva come i lavori con una maggiore retribuzione siano quello del *Manager* e del *Research Director* con un salario in media di molto superiore rispetto a tutti gli altri lavori, inoltre queste due classi rappresentano solo il 12% circa del totale.

Viene proposta ora una visualizzazione grafica grazie all'utilizzo dei boxplot in base alle classi della variabile **JobRole** sempre in relazione alla variabile **MonthlyIncome**.

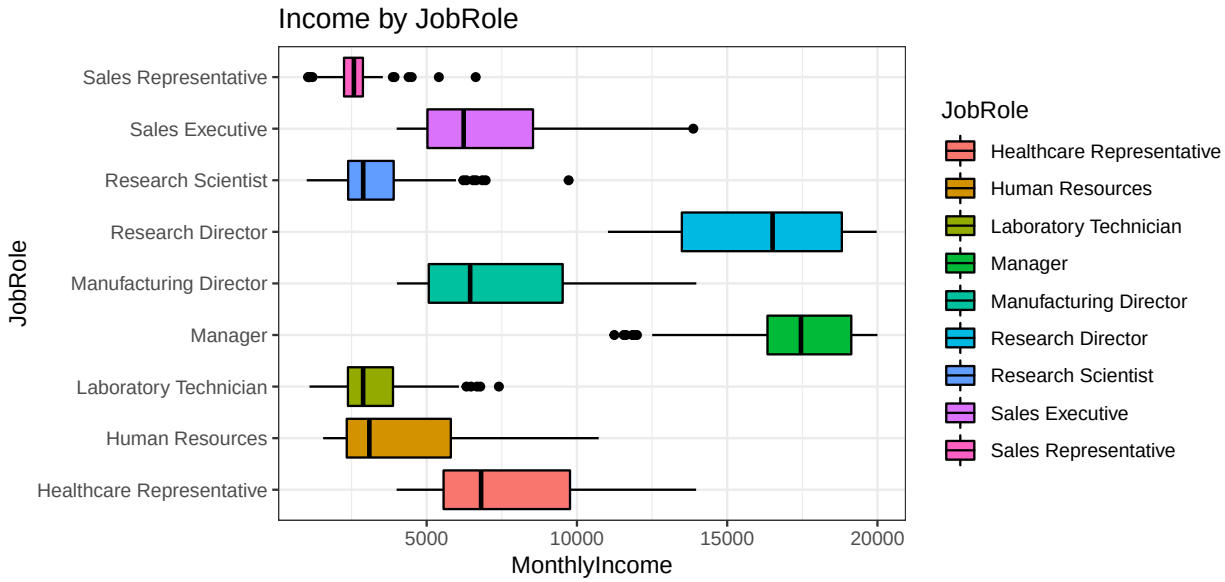


Figure 14

Dal grafico una si nota una varianza molto elevata per quanto riguarda i dipendenti che svolgono il ruolo di *Research Director*, quindi una forte variazione negli stipendi, si evince anche che non sono presenti outliers. Un altro lavoro ruolo che assume una distribuzione particolare è *Sales Representative*: sono infatti i dipendenti che guadagnano di meno e presentano anche una varianza minore, quindi gli stipendi saranno tutti molto simili eccetto che per qualche outliers sia nella parte inferiore sia nella parte superiore della distribuzione. Per quanto riguarda i *Manager*, che come si è visto sono coloro che percepiscono uno stipendio mediamente maggiore rispetto agli altri dipendenti, si osservano outliers nella parte inferiore della distribuzione, forse perchè più giovani rispetto agli altri colleghi.

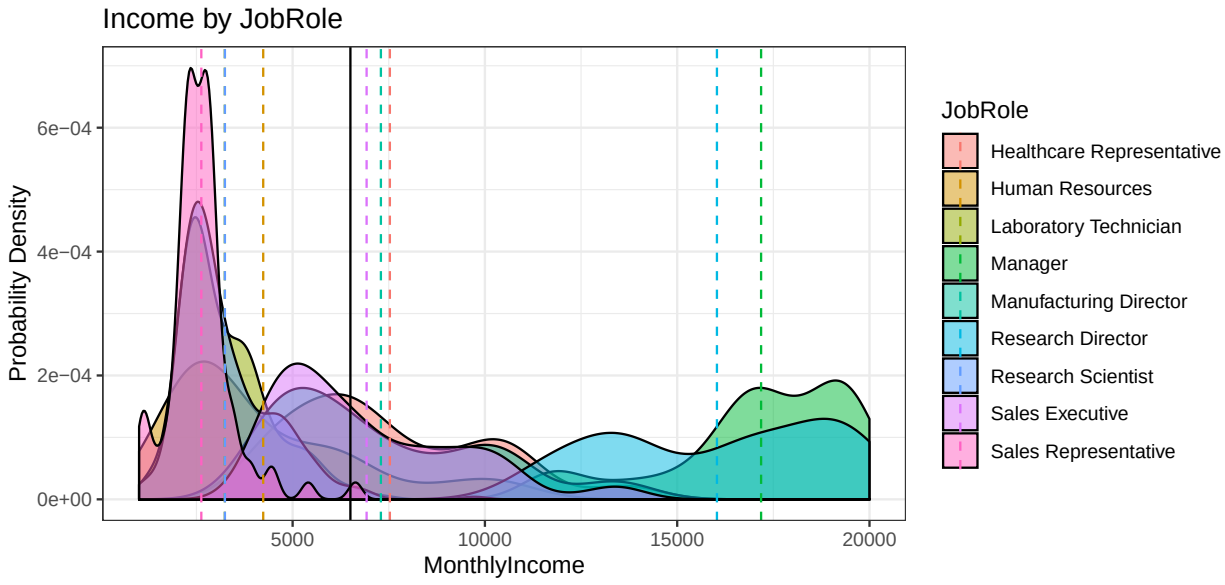


Figure 15

Nel grafico precedente vengono visualizzate le distribuzioni di probabilità e si evince, tra i vari **JobRole**, una varianza nei gruppi molto elevate per tutti i gruppi. Ciò è dovuto fortemente al grosso numero di outliers presenti all'interno di ciascun gruppo. Per quanto riguarda invece la varianza tra i gruppi si può osservare come questa risulti molto varia: la media della maggior parte dei gruppi è vicina alla linea nera rappresentante la media totale della variabile **JobRole**, per quanto riguarda i due lavori più pagati visti precendentemente la loro media è molto lontana.

Table 24: Analysis of Variance Table

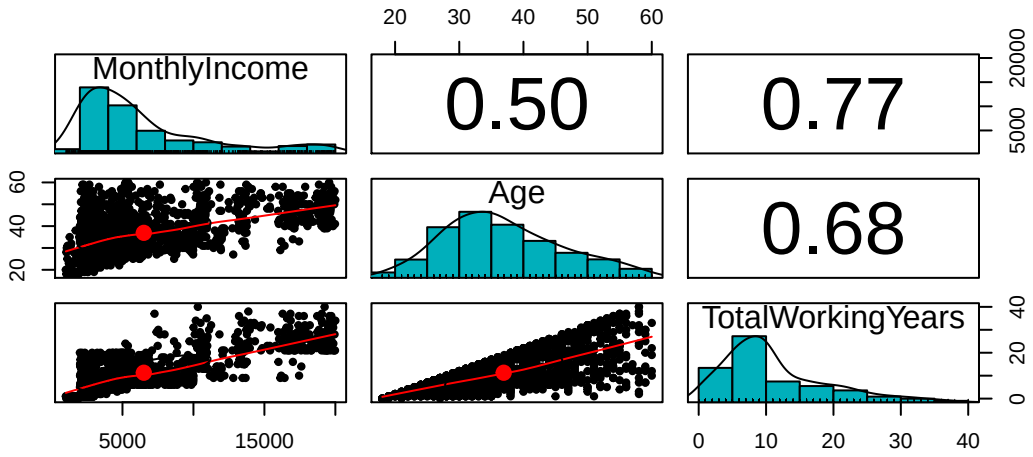
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
JobRole	8	2.657e+10	3.321e+09	810.2	0
Residuals	1,461	5.989e+09	4,099,377	NA	NA

Osservando i risultati dell'ANOVA si vede come il p-value sia praticamente uguale a zero, ovvero il valore F trovato è di molto superiore a quello critico consentito, quindi può essere rifiutata l'ipotesi nulla di uguaglianza tra le medie dei campioni e si può affermare che con i dati presenti c'è una significativa differenza tra le medie dei salari in base al tipo di lavoro.

3. Regressione Lineare

In questo ultimo capitolo verranno utilizzatate due regressioni lineari per predirre il valore di una data variabile quantitativa Y (**MonthlyIncome**), ovvero variabile dipendente, attraverso una o più variabili X definite regressori.
Di seguito, verranno proposte due tipi di regressioni: la prima lineare semplice con una sola variabile esplicativa (**TotalWorkingYears**), nel secondo caso verrà proposta una regressione lineare multipla con due variabili esplicative (**TotalWorkingYears** e **Age**). L'obbiettivo è quello di prevedere il salario, infatti è ragionevole pensare che quest'ultimo possa essere ben definito in base all'età di ogni singolo dipendente e la totalità degli anni lavorativi di questi ultimi.

Come prima cosa è utile analizzare la tabella di correlazione delle variabili che verranno usate nella regressione in modo da poter capire a priori se il modello possa essere affetto da multicollinearità.



Questo grafico è molto utile perche fornisce molte informazioni: sono rappresentate le correlazioni tra le variabili, le distribuzioni delle tre variabili con i relatiivi istogrammi e le curve di densità di probabilità. Si nota infatti come tutte le variabili non risultino avere una distribuzione normale e, al contrario, presentano tutte e tre assimmetria positiva.
Si evince infatti come le correlazioni siano buone (ottima per quanto riguarda **MonthlyIncome** e **TotalWorkingYears**) dato che sono tutte superiori al 0.5.
Si può anche affermare che la probabilità di osservare multicollinearità all'interno di un modello di regressione con le seguenti variabili sia bassa dato che la loro correlazione non ragginge livelli allarmanti superiori al 90%.

3.1 Regressione lineare semplice

Viene ora proposta la regressione lineare semplice MonthlyIncome ~ TotalWorkingYears e vengono presentati i relativi output di regressione:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1,228	137.3	8.944	1.108e-18
TotalWorkingYears	467.7	10.02	46.67	2.729e-292

Table 26: Fitting linear model: MonthlyIncome ~ TotalWorkingYears

Observations	Residual Std. Error	R^2	Adjusted R^2
1470	2988	0.5974	0.5971

Dall'output si evince un buon, seppur non elevatissimo R^2 pari a 0.5974 che sta ad indicare il fitting del modello, ovvero in che proporzione il modello sia in grado di spiegare i dati in possesso.
Fondamentale per il modello è che i P-value delle variabili prese in esame siano inferiori al livello di significatività scelto ($\alpha = 0.05$): i p-value sono infatti di gran lunga inferiori al livello prefissato e può quindi rifiutata l'ipotesi nulla di non significatività delle variabili per il modello, affermando che vi è una relazione tra le due variabili che sarà difficilmente dovuta al caso.
Considerando invece la colonna *Estimate*, che sta ad indicare i coefficienti veri e propri della retta di regressione, si vede come, per assurdo, un dipendente con 0 anni di lavoro alle spalle percepisca uno stipendio pari a 1227.94 che verrà incrementato di 476.66 dollari per ogni anno lavorativo aggiuntivo.

Nel grafico seguente viene presentata la visualizzazione grafica della retta di regressione.

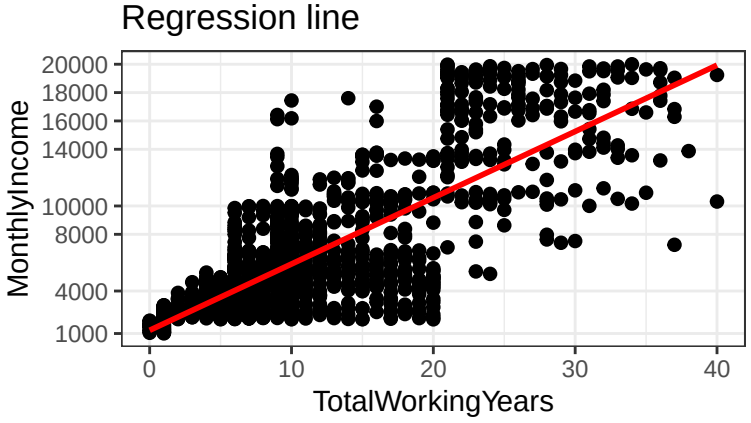


Figure 16

Successivamente, come si può vedere dal grafico seguente, è utile capire come si distribuiscono i residui di regressione in relazione ai valori (TotalWorkingYears) predetti dal modello.

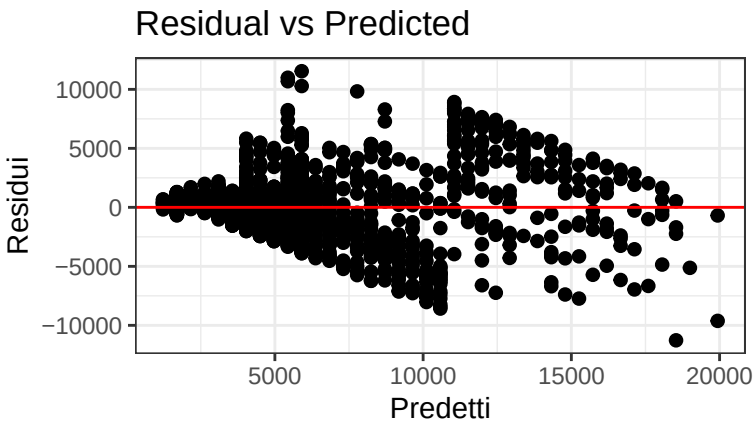


Figure 17

I residui dovrebbero disporsi casualmente intorno alla linea rossa che sta ad indicare lo scarto che si vorrebbe ottenere tra valori osservati e valori predetti dal modello (si vuole infatti che lo scarto sia minimo, quindi uguale a zero).
Da questo grafico si nota una certa dispersione crescente dei punti che tendono ad allargarsi (fino a poco più di 10.000 dollari) per poi contrarsi assumendo comunque un ordine sparso nella parte destra del grafico.
Questa forma implica un alto valore dei residui di regressione, che comporta una varianza non costante all'interno del modello, la quale potrebbe essere sintomo di eteroschedsticità dei residui, si può inoltre notare la presenza di alcuni outliers che se rimossi potrebbero migliorare i valori del modello.

Il passo successivo è quello di valutare se i residui del modello presentano una distribuzione normale in modo da poter applicare eventuali intervalli di confidenza.
Nel grafico seguente vengono proposti il QQplot e l'istogramma relativo alla distribuzione dei residui.

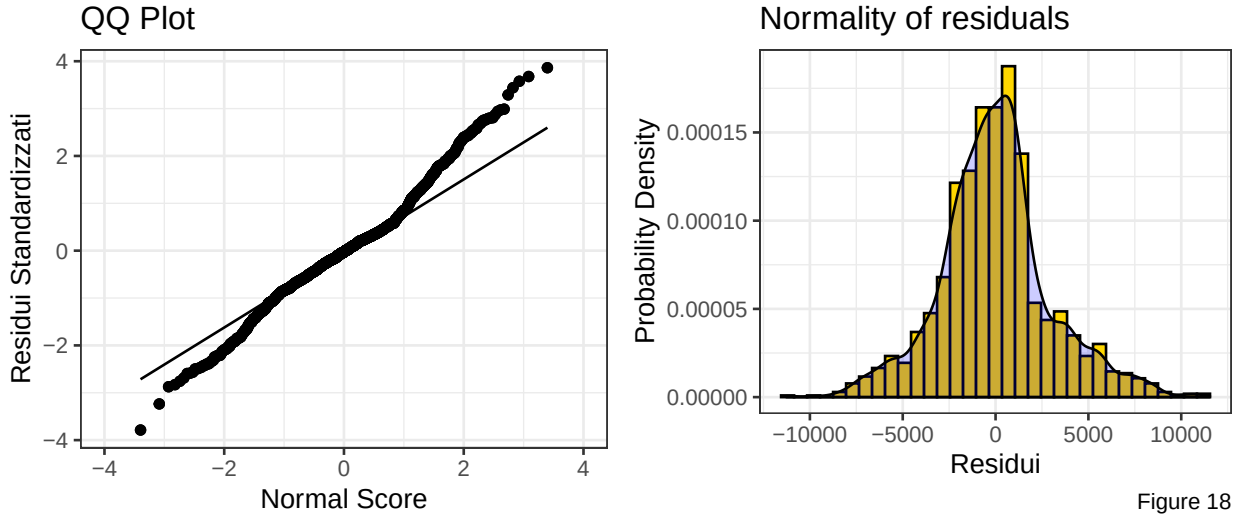


Figure 18

Per poter affermare che i residui del modello siano normali occorre che i punti del QQplot si distribuiscano il più possibile sulla retta, cosa che non accade in questo modello di regressione. Anche l'istogramma dei residui riporta code pesanti e una campana con una punta più stretta, infatti la curva di densità di probabilità è leptocurtica dato che il test di curtosi ha dato un valore prossimo a 4, si può quindi affermare che i residui del modello non presentano una distribuzione normale.

3.2 Regressione lineare multipla

Quest'ultima sezione è molto simile alla precedente, ma presenta un regressore in più, vale a dire la variabile **Age**. Si prosegue quindi con il summary relativo al nuovo modello di regressione multipla.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1,978	352.4	5.614	2.365e-08
TotalWorkingYears	489.1	13.65	35.82	1.838e-202
Age	-26.87	11.63	-2.311	0.02097

Table 28: Fitting linear model: MonthlyIncome ~ TotalWorkingYears + Age

Observations	Residual Std. Error	R^2	Adjusted R^2
1470	2984	0.5988	0.5983

Si intuisce come l'aggiunta della variabile Age al modello renda praticamente invariato il valore dell' R^2 normale e di quello aggiustato, risulta significativa anche la nuova variabile con un livello di significatività sempre dello 0.05.
Da notare sono i cambiamenti relativi ai parametri della retta di regressione: l'intercetta è passata da circa 1200 a 1900, aumenta di poco il parametro relativo a TotalWorkingYears mentre sembra che la variabile Age penalizzi la predizione finale del salario, infatti per ogni anno in piu di ogni dipendente si ha una perdita di salario predetto di circa 27$.

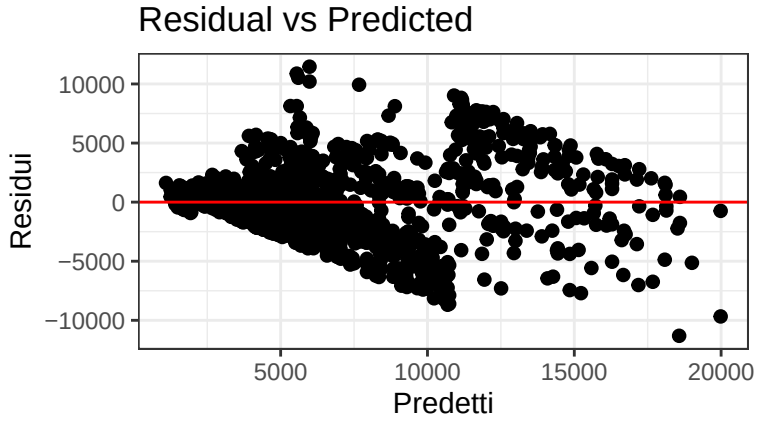


Figure 19

Si può affermare che anche il grafico dei residui del nuovo modello sia molto simile a quello relativo alla regressione precedente: anche in questo caso i residui tendono ad essere concentrati maggiormente per valori più bassi di salario mantenendo comunque questa dispersione, sempre maggiore, intorno alla linea rossa, tendendo poi a contrarsi di di nuovo.

Da i due grafici sottostanti si può osservare che anche la normalità dei residui non sembra essere migliorata dopo l'aggiunta della variabile **Age** al modello.

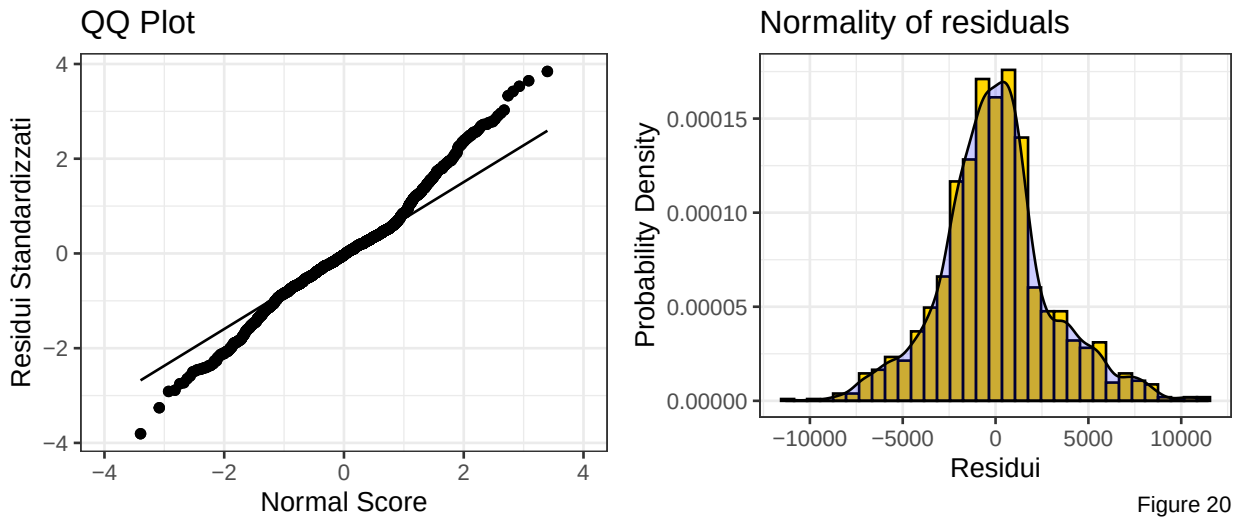


Figure 20

L'ultimo passo è quello di valutare se, e in che quantita, il modello è affetto da multicollinearità. In precedenza si è visto come la correlazione tra le variabili Age e TotalworkingYears fosse abbastanza alta ma non tanto da far pensare a multicollinearità.
Nelle due tabelle sottostanti vengono presentati il VIF, indice che ci pertmette di valutare il grado di multicollinearità: un valore superiore a 10 è da considerarsi preoccupante, tuttavia non esiste un valore universale che indica quando è presente multicollinearità.
Nella seconda tabella viene presentata la radice quadrata del VIF che sta ad indicare quanto è più grande l'errore standard, rispetto ad un altro regressore, nel caso in cui questo regressore fosse incorrelato con le altre variabili esplicative. In questo caso il valore limite da non superare viene posto uguale a 2.

TotalWorkingYears	Age
1.862	1.862

TotalWorkingYears	Age
FALSE	FALSE

Si evince che i VIF sono particolarmente bassi e nessuna delle due variabili ha un $\sqrt{VIF}$ maggiore di due dato che gli output sono stati entrambi *FALSE* per le due variabili.
Si può dunque affermare che questo modello di regressione non presenta multicollinearità.