Consegna prova d'esame Programmazione e Analisi Dati

Arena (544907)[*] Bergonzini (560680)[†] Biondi (547237)[‡] Ricca (641031)[§]

0.1 Indice:

[*] a.arena11@studenti.unipi.it
[†] a.bergonzini1@studenti.unipi.it
[‡] r.biondi@studenti.unipi.it
[§] a.ricca8@studenti.unipi.it

0.1.1 *Introduzione*

Con il seguente studio il nostro gruppo si è posto come obiettivo svolgere uno studio di tipo **qualitativo** sul dataset HR14. In particolare, l'approccio scelto è di tipo top-down.

Come si vedrà in seguito, questo tipo di approccio ci conduce da un tipo di grafico meramente quantitativo a grafici di analisi qualitativa. Abbiamo iniziato lo studio partendo da una sezione dedicata ad alcuni calcoli statistico-descrittivi e alcuni grafici generici sul dataset. Successivamente abbiamo diviso lo studio in tre macro aree: la prima sulle piattaforme di recruiting, la seconda sull'apporto salariale degli impiegati, la terza sulle correlazioni presenti nel dataset, più una sezione extra con alcuni grafici aggiuntivi. Nella sezione riguardante le correlazioni, oltre ai grafici, è presente un'analisi sull'andamento delle distrubizoni.

Infine alcune considerazioni conclusive sul lavoro svolto.

0.2 *Sezione Statistica descrittiva*

Di seguito una tabella che mostra i dati statistici (media, deviazione standard, quartili, massimo e minimo) delle colonne del dataset che presentano dati numerici.

	Salary	**Age**	**EngagementSurvey**	**Absences**	**AgeTerm**	**AgeofHire**
count	311.00	311.00	311.00	311.00	104.00	311.00
mean	69020.68	43.41	4.11	10.24	37.65	34.10
std	25156.64	8.87	0.79	5.85	9.88	8.91
min	45046.00	30.00	1.12	1.00	23.00	19.00
25%	55501.50	36.00	3.69	5.00	29.75	28.00
50%	62810.00	42.00	4.28	10.00	35.50	32.00
75%	72036.00	49.00	4.70	15.00	43.25	39.00
max	250000.00	71.00	5.00	20.00	64.00	63.00

La seguente tabella mostra la mediana di ciascuna colonna con valori numerici.

Salary	62810.00
Age	42.00
EmpSatisfaction	4.00
EngagementSurvey	4.28
Absences	10.00
DaysLateLast30	0.00
SpecialProjectsCount	0.00
AgeTerm	35.50
AgeofHire	32.00

Di seguito invece, in ordine, i dati dei primi cinque dipendenti con stipendio maggiore e successivamente dei cinque impiegati più anziani.

Salario

Employee_Name	Salary	Position	Department	Age
King, Janet	250000	President & CEO	Executive Office	68
Zamora, Jennifer	220450	CIO	IT/IS	43
Houlihan, Debra	180000	Director of Sales	Sales	56
Foss, Jason	178000	IT Director	IT/IS	42
Corleone, Vito	170500	Director of Operations	Production	39

Età

Employee_Name	Salary	Position	Department	Age
Chace, Beatrice	61656	Production Technician I	Production	71
Daniele, Ann	85028	Sr. Network Engineer	IT/IS	70
King, Janet	250000	President & CEO	Executive Office	68
Ren, Kylo	61809	Area Sales Manager	Sales	68
Biden, Lowan M	64919	Production Technician I	Production	64

Per concludere la sezione sulla statistisca descrittiva riportiamo una tabella mostrante lo ZScore di ogni record dei valori numerici del dataset.

$Z = \frac{x-\overline{x}}{s}$

Lo ZScore mostra il valore della deviazione standard, superiore o inferiore alla media: più è vicino allo 0, più il valore si avvicina alla media.

Salary	Age	EngagementSurvey	Absences	AgeofHire
-0.26	-0.50	0.62	-1.58	-0.69
1.41	0.41	1.08	1.16	0.66
-0.16	-1.06	-1.38	-1.24	-1.25
-0.16	-1.06	0.93	0.81	-1.58
-0.72	-1.18	1.13	-1.41	-1.36

Grafico 1: Rapporto uomini-donne all'interno dell'azienda Nel grafico viene analizzato il rapporto tra dipendenti di genere maschile e dipendenti di genere femminile; si evince una leggera maggioranza femminile all'interno del personale.

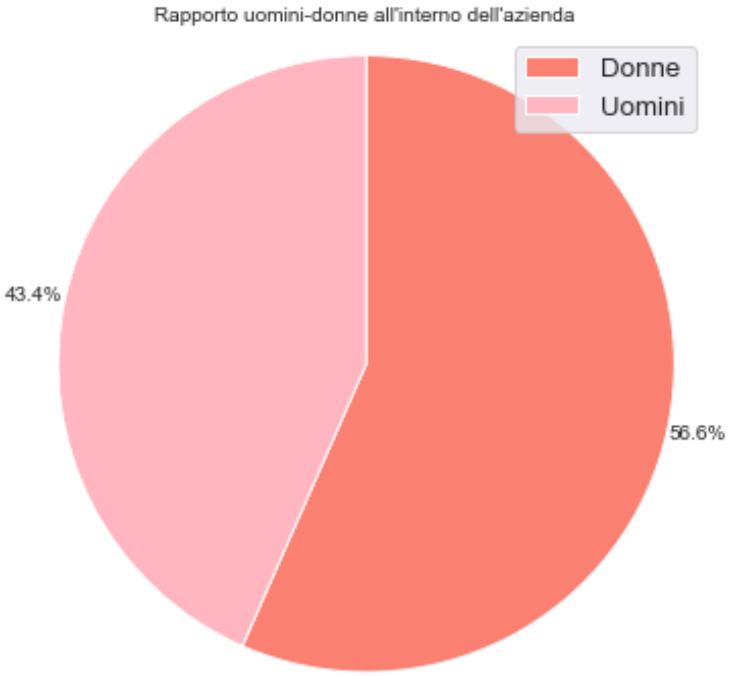
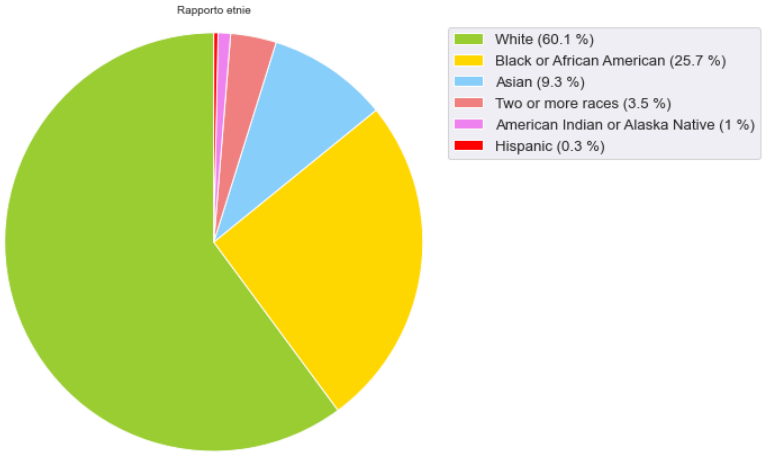


Grafico 2: Rapporto etnie Il seguente grafico analizza il rapporto tra le differenti etnie. Mentre la maggior parte degli impiegati è di etnia caucasica, notiamo come persone di colore e afroamericani ed asiatici rappresentino le altre due grandi maggioranze etniche presenti; il restante 4.8% è occupato da (in ordine): due o più etnie, nativi americani e ispanici.



---

1 *Sezione Piattaforme di assunzione*

Nella segeunte sezione ci siamo impegnati ad analizzare nello specifico come la piattaforma di recruiting, dalla quale gli impiegati sono stati selezionati, potesse influire nella vita lavorativa degli stessi, partendo dalle semplici percentuali di assunzione (Grafico 3). Abbiamo cercato di rispondere a domande quali l'esistenza di una relazione tra l'assunzione tramite queste piattaforme e le *prestazioni* dei singoli dipendenti, *l'età al momento dell'assunzione* o il *dipartimento* di appartenenza.

Grafico 3: Percentuali piattaforme di assunzione Il seguente grafico analizza il rapporto tra le differenti piattaforme di assunzione, attraverso le quali gli impiegati sono stati assunti all'interno dell'azienda. Seppure la maggioranza degli impiegati venga dalle piattaforme web-social (LinkedIn - Indeed) e oltre l'80% degli stessi sia stato assunto attraverso canali online; è altresì interessante notare come ben il 10% degli impiegati sia stato assunto *analogicamente* su consiglio di lavoratori già presenti. Nota di merito per il programma di *diversity job fair* il quale prevede assunzioni (9.3%) per persone appartenenti a categorie protette, con disabilità e di origine straniera.

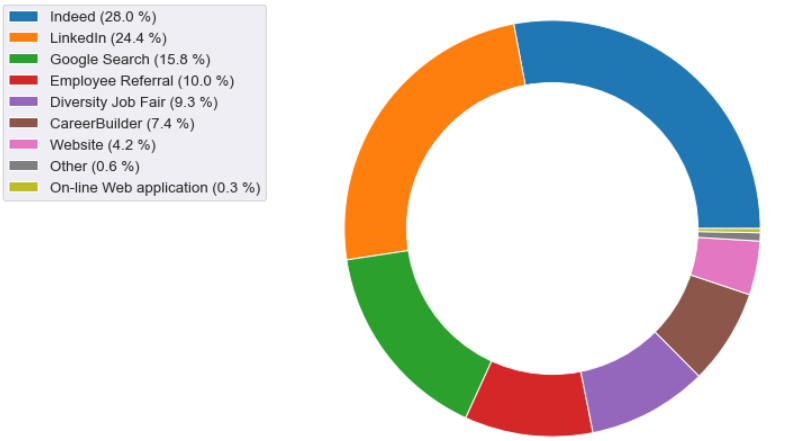

Indeed (28.0 %)
LinkedIn (24.4 %)
Google Search (15.8 %)
Employee Referral (10.0 %)
Diversity Job Fair (9.3 %)
CareerBuilder (7.4 %)
Website (4.2 %)
Other (0.6 %)
On-line Web application (0.3 %)

Grafico 4: Rapporto tra il Performance Score e la piattaforma di recruiting Questo grafico analizza il rapporto tra le differenti piattaforme di assunzione e il *performance score* medio. Si evince come il punteggio si attesti in valori compresi tra 2.5 e 3.0 con i soli impiegati derivanti da dall' *employee referral* che superano questa media.

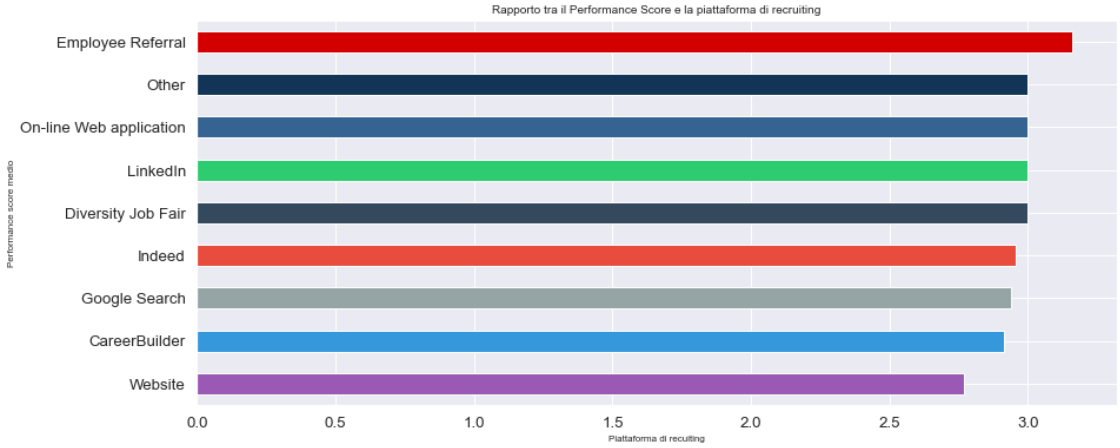
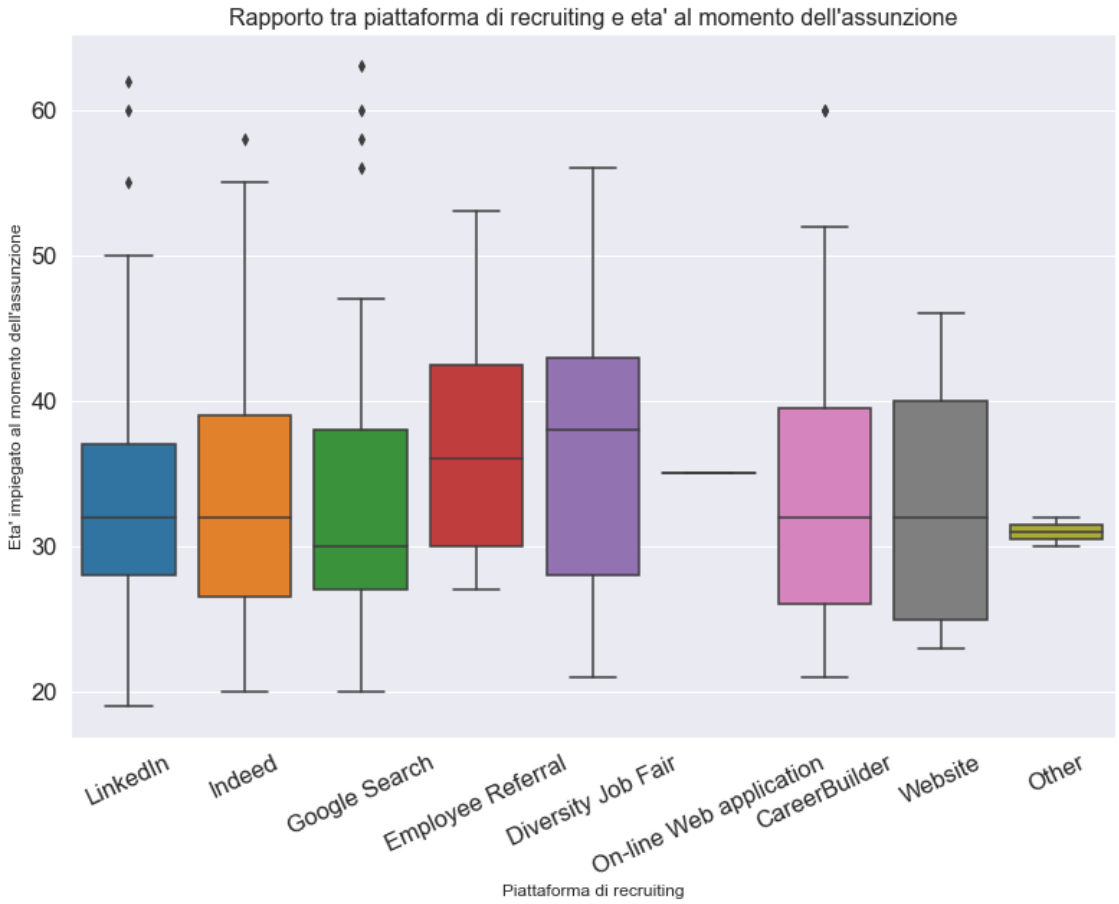
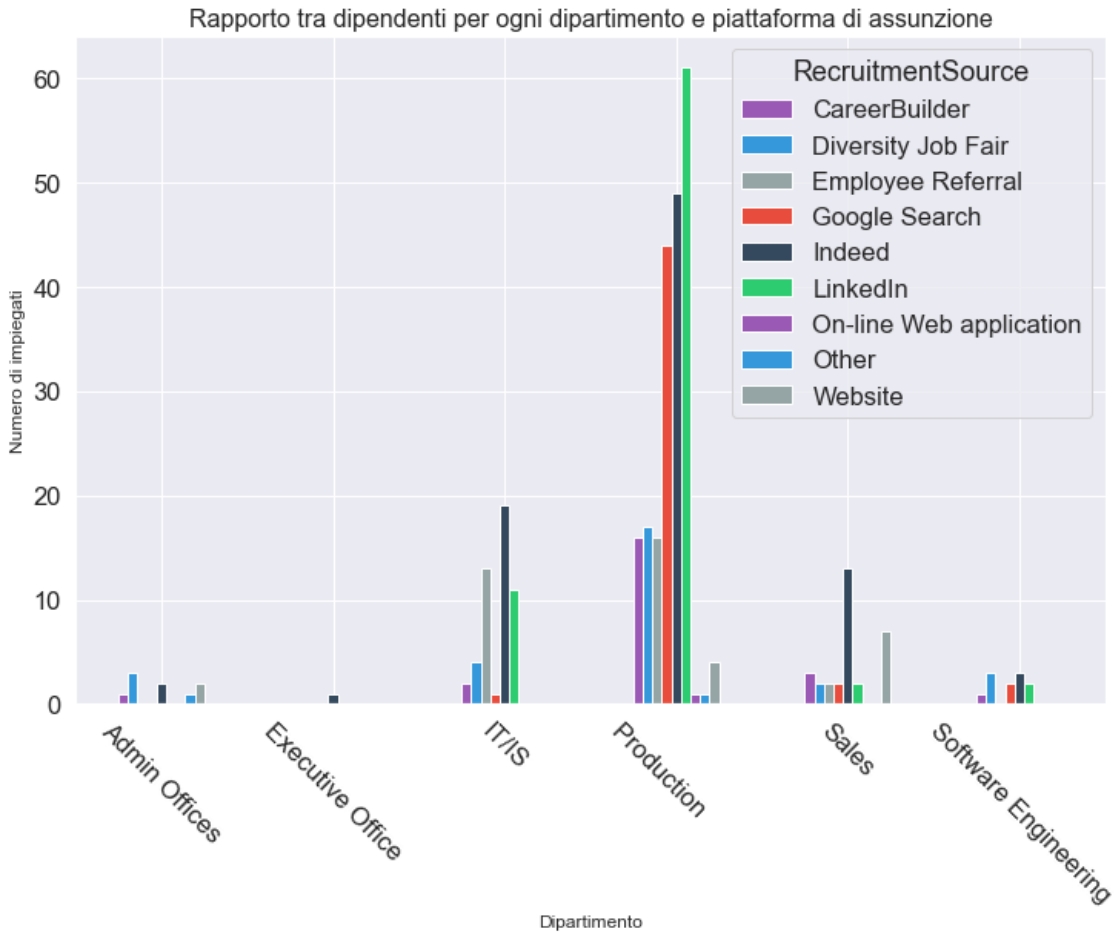


Grafico 5: Rapporto tra piattaforma di recruiting e eta' al momento dell'assunzione In questo boxpolot analiziamo il rapporto tra l'età media di assunzione e la piattaforma di recruiting. Abbiamo voluto mettere in mostra l'età dei dipendenti provenienti da ogni piattaforma lavorarativa. Molto interessanti sono le piattaforme *Google Search* e *Online Web Application* i cui dati, a differenza delle altre scatole, sono asimmetrici con la linea di mediana leggermente scostata dal centro. Ancora, mentre in *Google Search* possiamo osservare ben 3 outliers, è altrettanto interessante notare come l'*Online Web Application* e la categoria *Others* abbiano praticamente un IQR pari a 0.



Piattaforma di recruiting	Numero di impiegati
Indeed	87
LinkedIn	76
Google Search	49
Employee Referral	31
Diversity Job Fair	29
CareerBuilder	23
Website	13
Other	2
On-line Web application	1

Grafico 6: Rapporto tra dipendenti per ogni dipartimento e piattaforma di assunzione Nel seguente grafico abbiamo messo in relazione i dati provenienti dal *dipartimento* e dalla *piattaforma di recruiting*. E' stata creata una crosstab per calcolare il numero di persone provenienti da un determinato dipartimento, le quali sono state assunte tramite le varie piattaforme. Il dipartimento con il più alto numero di persone è quello della *production*, con il maggior numero di dipendenti assunto via LinkedIn. Inoltre nel dipartimento Executive Office l'unico picco, proveniente da Indeed, è rappresentato dalla CEO dell'azienda.



---

2 *Sezione Salario*

Qui ci occuperemo di indagare diversi aspetti riguardanti il contributo salariale degli impiegati, cercando di dimostrare se questo sia effettivamente legato alle singole *posizioni lavorative*, al *dipartimento*, al *sesso*, alla quantità di *progetti speciali* svolti e infine all'assunzione tramite progetti speciali come il *diversity job fair*.

Grafico 7: Relazione tra la posizione lavorativa e il guadagno medio Nel primo grafico della sezione si mostra la relazione tra la posizione lavorativa e il loro guadagno medio. Osserviamo come la maggior parte delle posizioni abbia un guadagno medio annuo inferiore ai 150000 dollari. Solo gli impiegati appartenenti alle prime sette categorie guadagnano annualmente più di questa cifra.

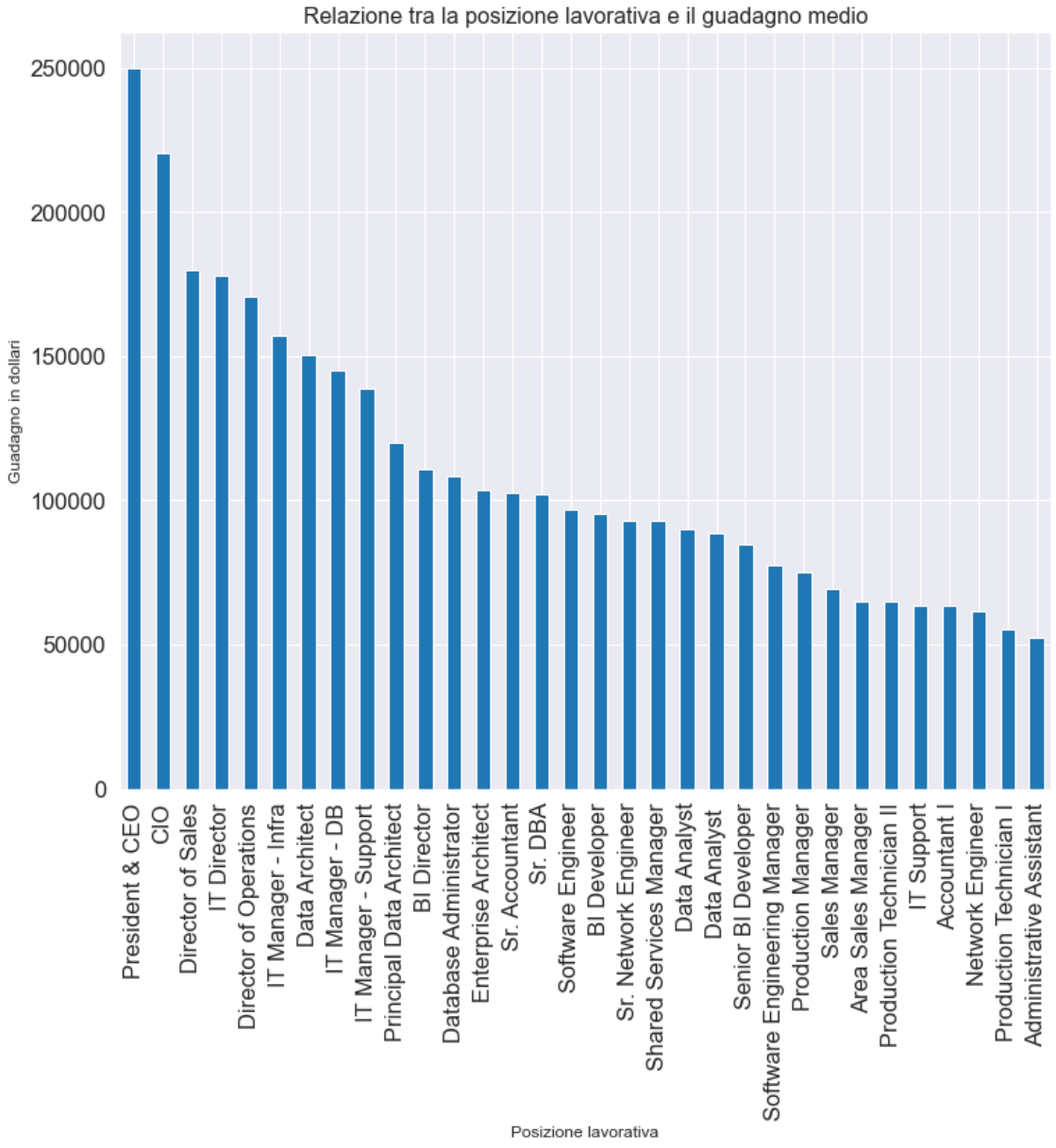


Grafico 8: Comparazione tra salario medio e dipartimento In questo pointplot mettiamo in relazione il *salario* medio con il *dipartimento* degli impiegati, rispetto al numero di membri di ognuno di quest'ultimo. Interessante osservare come vi sia una disparità di guadagno tra i dipartimenti con più impiegati rispetto al picco del lineplot raggiunto da una sola persona, ovvero dal CEO, a quota 250000 dollari.

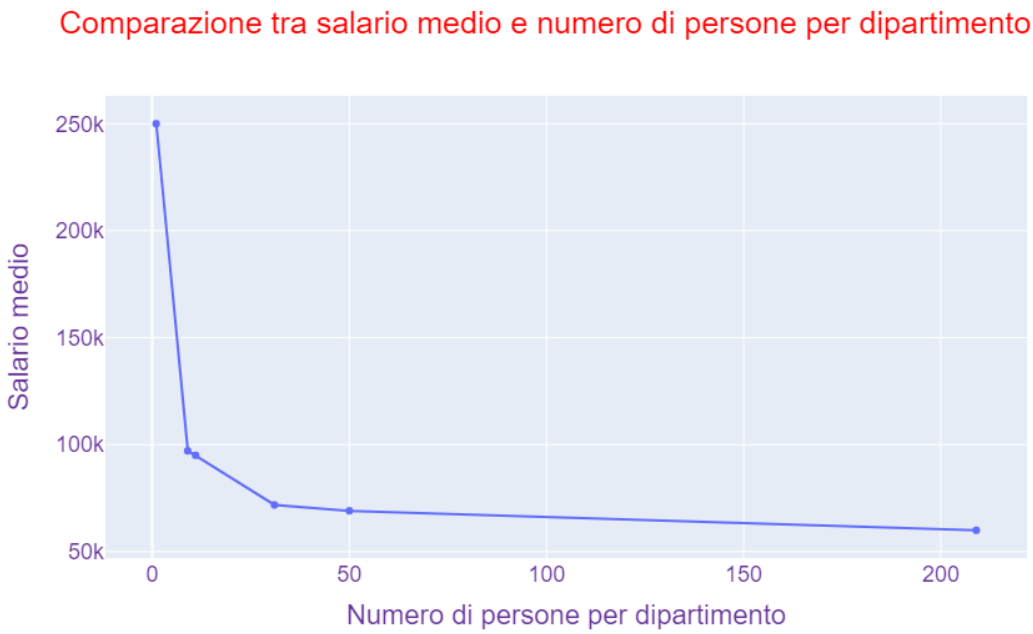


Grafico 9: Relazioni tra salario, dipartimento e numero di progetti speciali Dopo aver creato una crosstab tra il numero di *Special Projects*, i *Departmets* e il *Salary*, ed aver generato il grafico, notiamo come la maggior parte di progetti speciali vengano realizzati da persone impiegate nei campi dell'IT, dai programmatori software e dagli admin offices. Interessante che poche persone abbiano realizzato un solo progetto speciale ma che ogni dipartimento presenti qualcuno a non averne realizzati affatto.

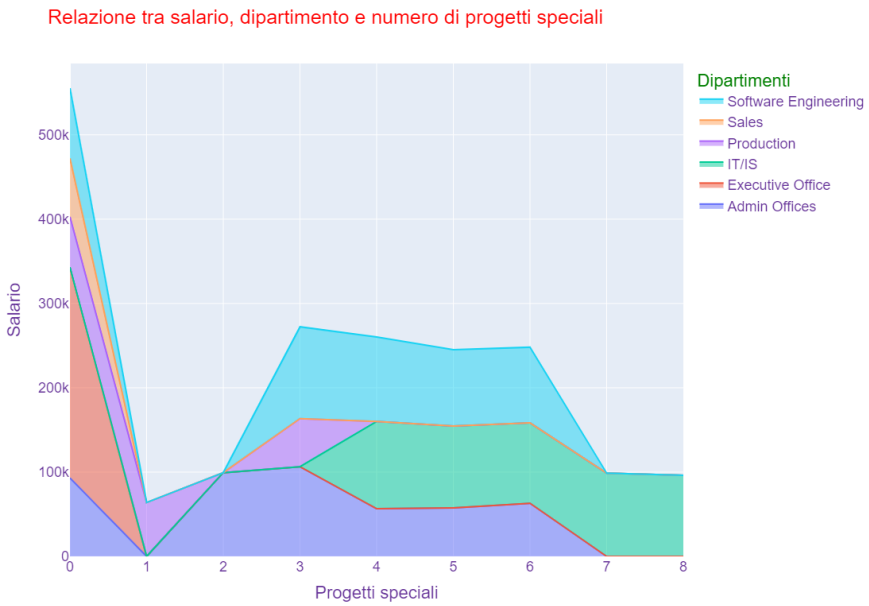


Grafico 10: Confronto tra dipartimento, sesso e salario In questo grafico possiamo osservare il rapporto tra il *salario medio* e il *dipartimento* in relazione con il *sesso* degli impiegati. Non sono presenti impiegati di sesso maschile nella categoria "Executive Office". Inoltre il salario più alto, 250.000 dollari, è assegnato a quella categoria in quanto rappresenta quella in cui è presente unicamente la CEO, la quale è una donna. Per il resto dei dipartimenti lo stipendio medio sembra essere abbastanza uniforme per entrambi i sessi.

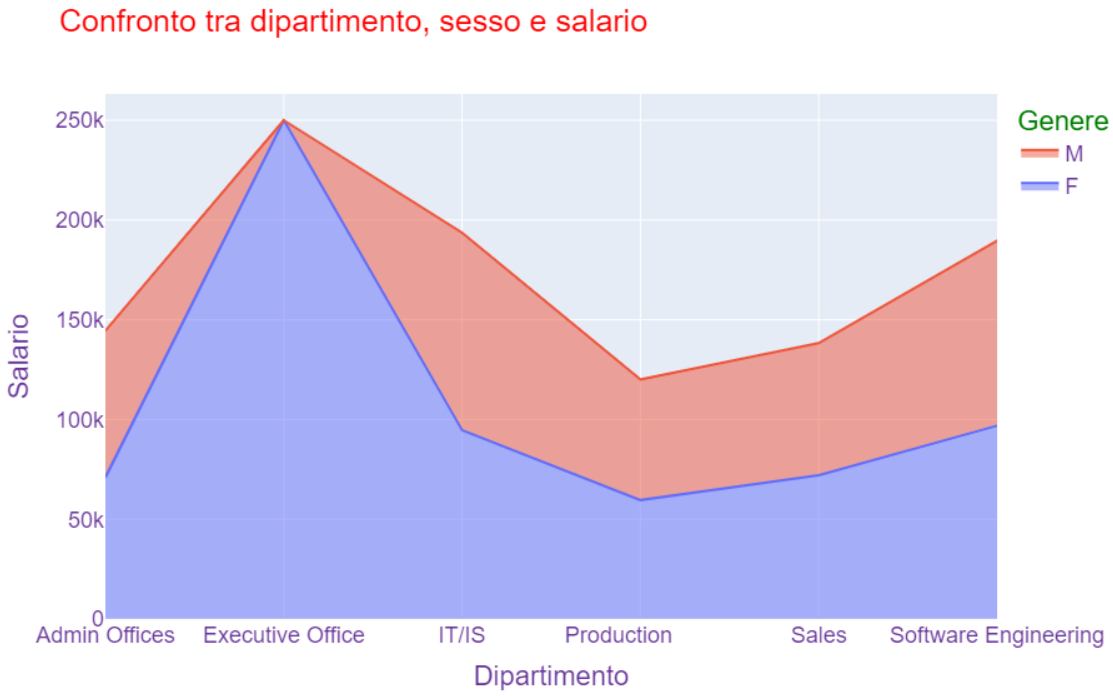
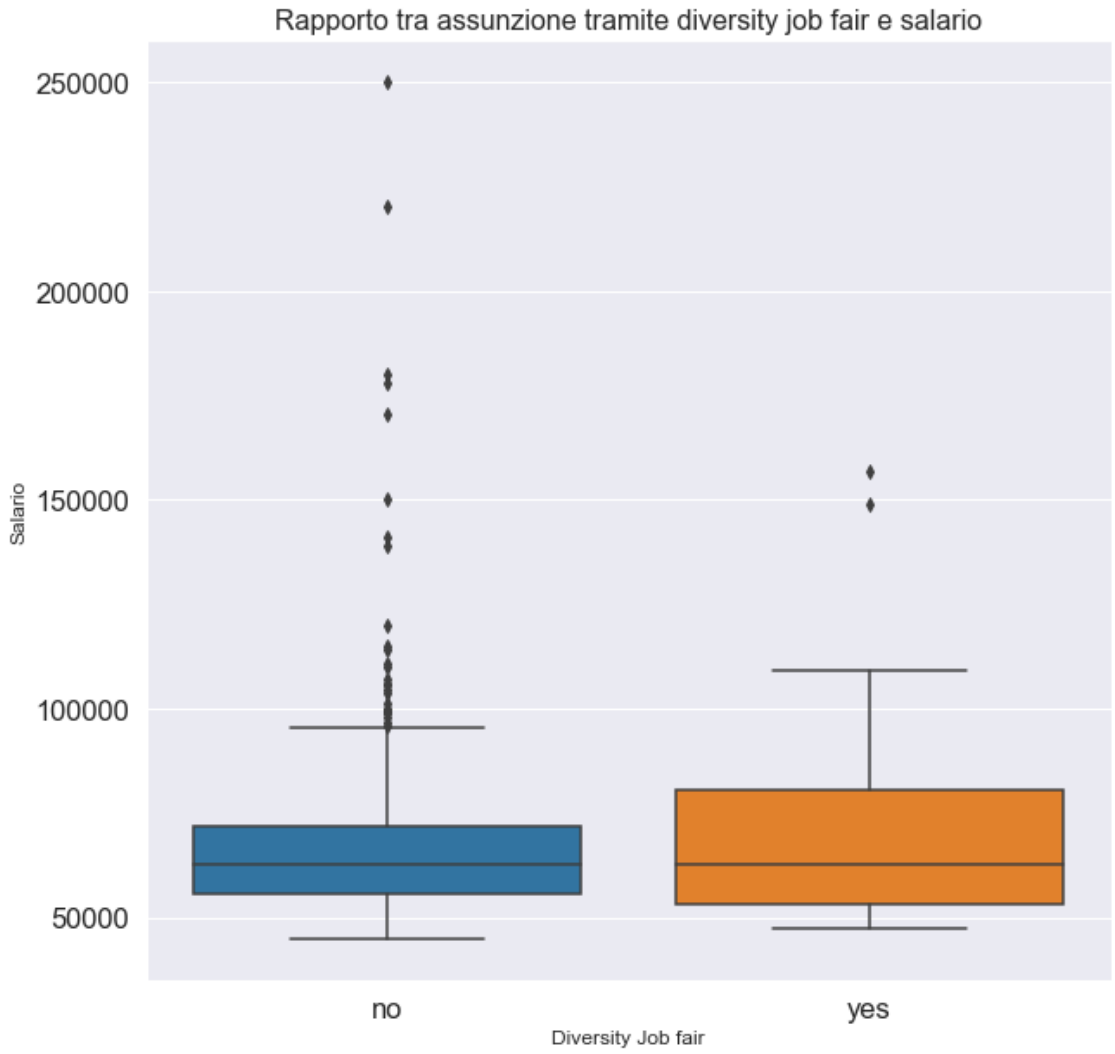
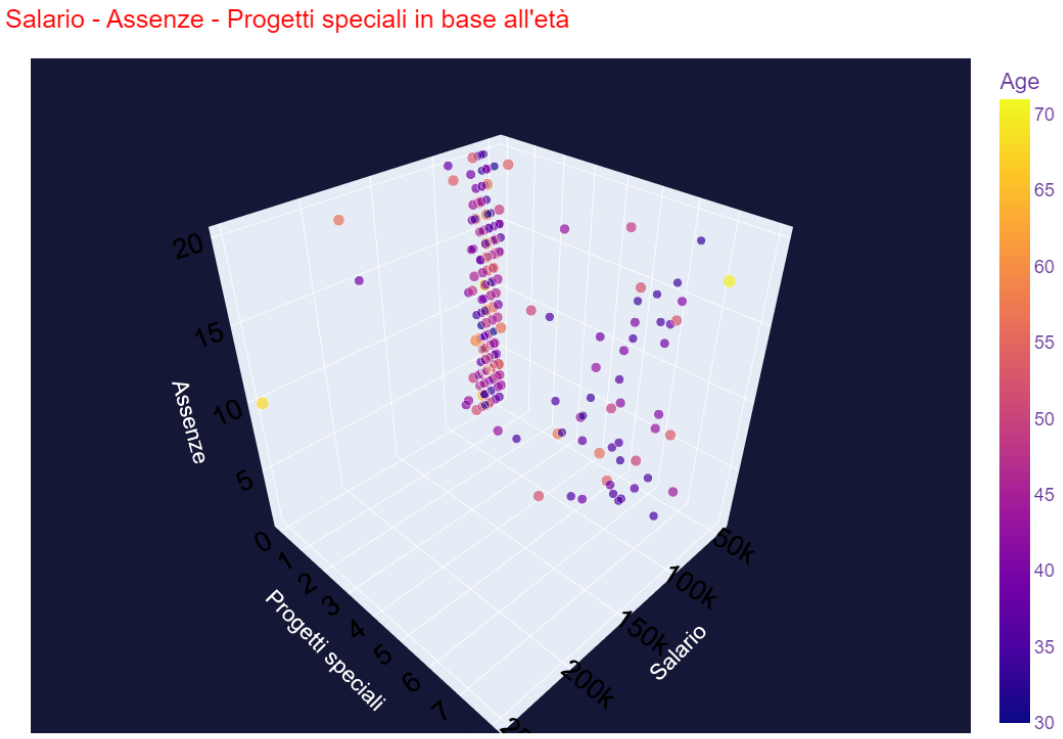


Grafico 11: Rapporto tra assunzione tramite diversity job fair e salario Il seguente boxplot mette in rapporto l'assunzione tramite il programma di *Diversity Job Fair* e il *salario*; osserviamo che la mediana presente nei due box è la medesima. Nonostante ciò, si osserva, tramite il baffo inferiore, che l'impiegato con il salario più basso assunto tramite questo programma ha comunque un salario maggiore rispetto all'impiegato non assunto tramite diversity job fair con il salario minore. Dagli outliers, osserviamo come la CEO non sia stata assunta tramite questo progetto. Solo due sono gli outliers relativi all'assuzione tramite il programma di diversity job fair.



Assunti attraverso il programma di *diversity job fair*?	
No	282
Si	29

Grafico 12: Rapporto tra salario, assenze e progetti speciali in base all'età dell'impiegato In questo grafico a bolle osserviamo il rapporto tra le variabili sopra citate. Ne emerge che la maggior parte dei dipendenti si posiziona ove il numero di progetti speciali è pari a 0, poichè appartenenti ai dipartamenti in cui questi non vengono svolti. Tra gli outlier relativi all'asse del salario spicca, come visto in precedenza, la CEO.



---

3 *Sezione Correlazioni*

Questa sezione si occupa di indagare se vi siano delle possibili correlazioni tra i dati presenti nel dataset, tramite i grafici e i calcoli degli indici di correlazione.

Grafico 13: Pairplot Al fine di mettere in correlazione tutti i dati numerici presenti all'interno del dataset abbiamo utilizzato un pairplot. Abbiamo realizzato una griglia 7x7 in cui possiamo notare come non vi sia nessuna correlazione significativa tra le varie combinazioni di coppie possibili.

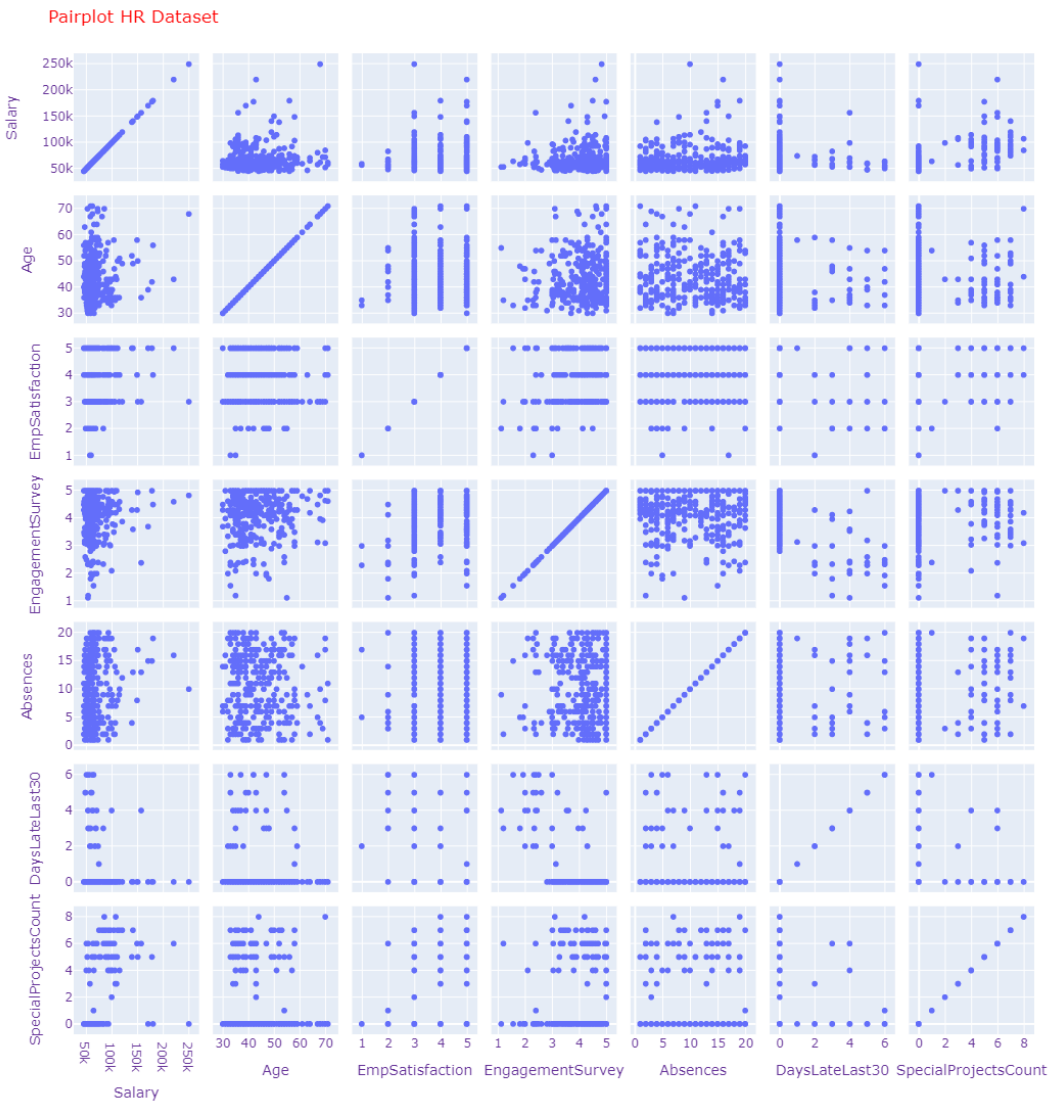


Grafico 14: Correlazione tra eta' e salario In questo scatterplot abbiamo messo in correlazione l'età degli impiegati e lo stipendio. Si evince come non vi sia una correlazione tra le due variabili. Altresì è da riportare che gli impiegati presi in considerazione sono esclusivamente quelli che tutt'ora risultano attivi nell'organigramma aziendale. Si deduce che la maggior parte degli impiegati ha un salario che si colloca sotto una determinata soglia.

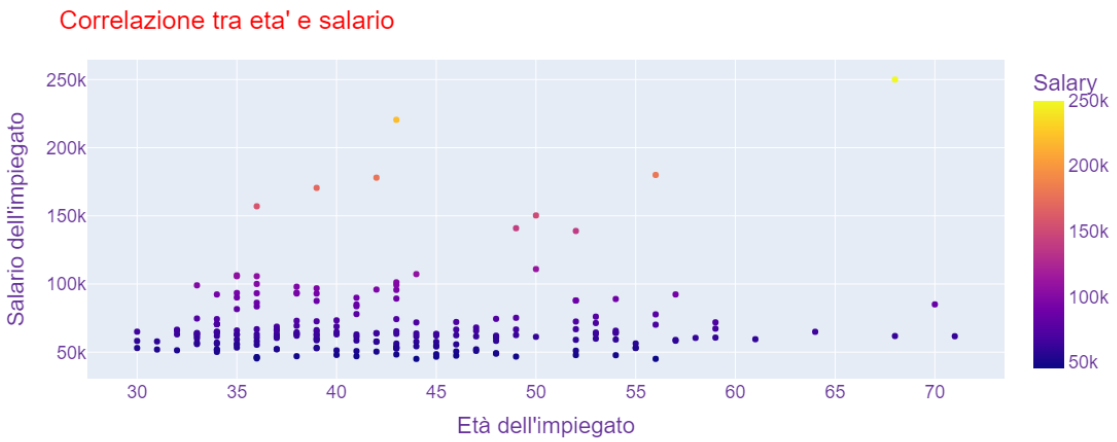


Grafico 15: Heatmap Di seguito diverse heatmap per analizzare correlazioni all'interno del dataset; quella subito sotto confronta tutte le variabili.

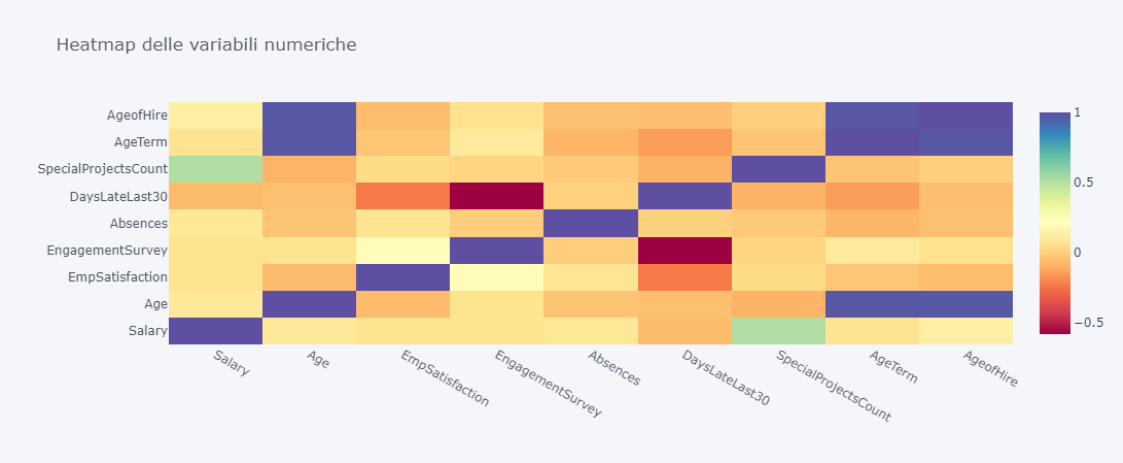


Grafico 15.1 : correlazione tra salario, eta', soddisfazione dell'impiegato e punteggio ottenuto al questionario dei tecnici di produzione del primo gruppo Per verificare la correlazione fra coppie di variabili in base alle posizioni specifiche del personale sono state realizzate alcune heatmap prendendo in considerazione le categorie lavorative più numerose. Qualora fosse presente una correlazione con un coefficente di Kendall abbastanza vicino a +1 o a -1 allora la correlazione è forte; mentre più è vicina allo 0 più la correlazione è assente.

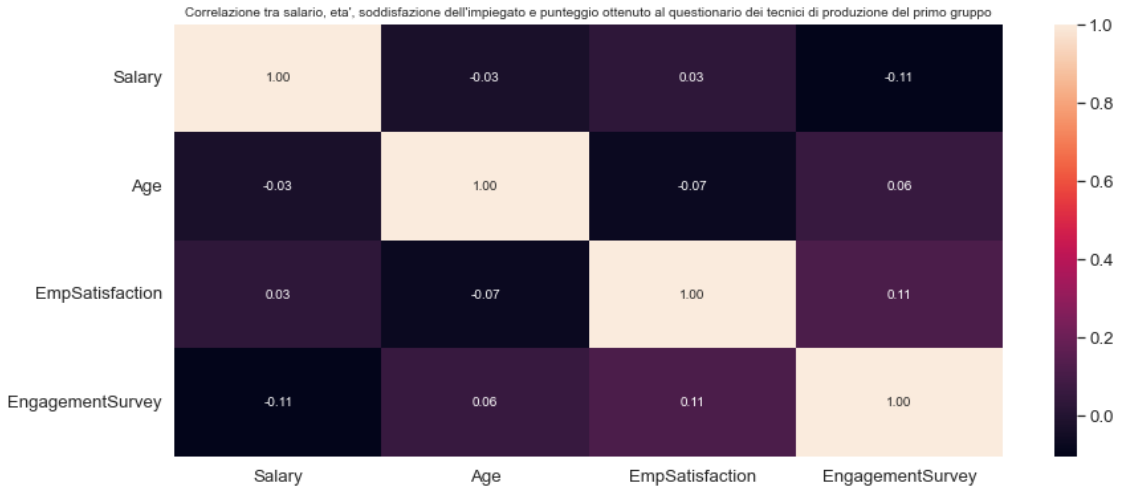


Grafico 15.2: Correlazione tra salario, eta', soddisfazione dell'impiegato e punteggio ottenuto al questionario dei tecnici di produzione del secondo gruppo

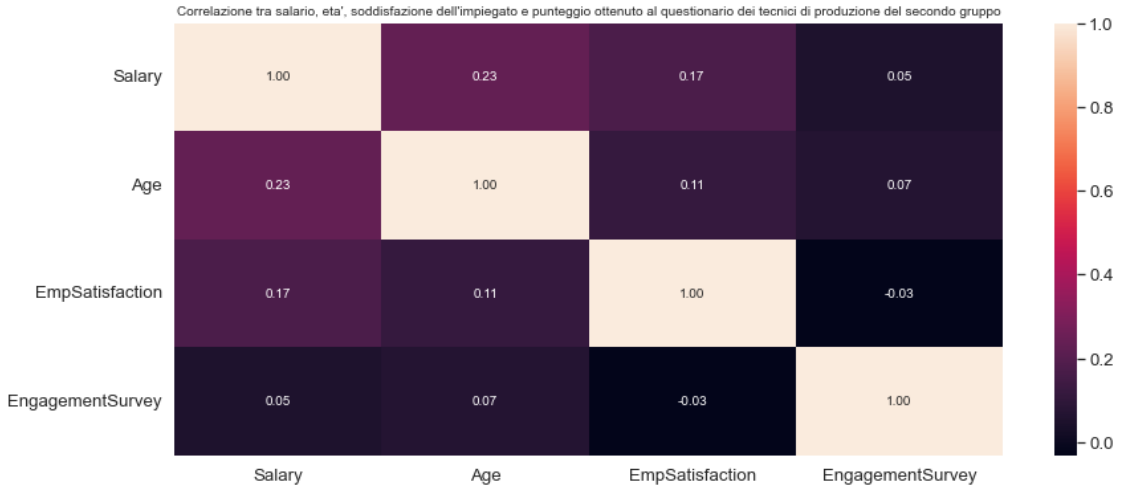


Grafico 15.3: Correlazione tra salario, eta', soddisfazione dell'impiegato e punteggio ottenuto al questionario dei manager dell'area vendite

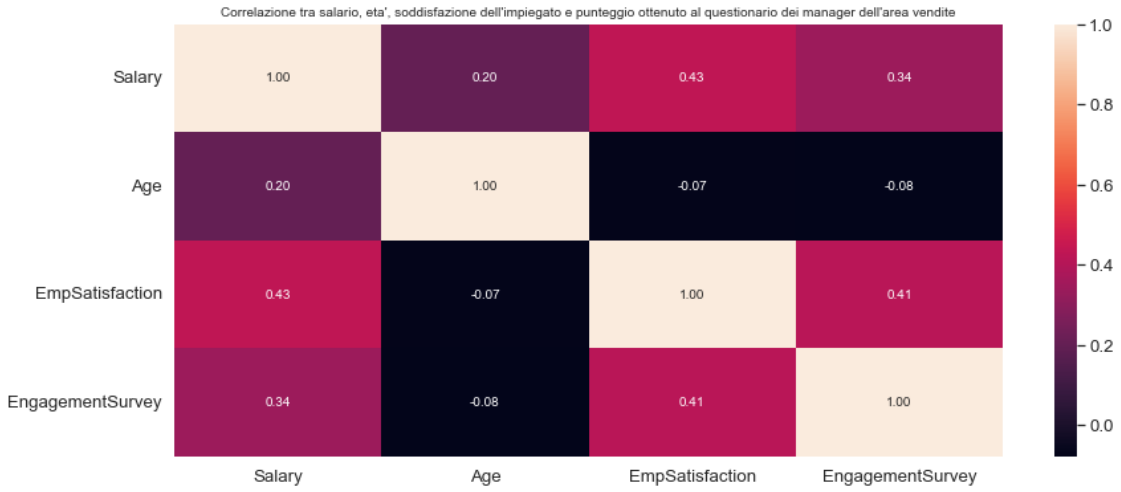


Grafico 15.4: Correlazione tra salario, eta', soddisfazione dell'impiegato e punteggio ottenuto al questionario dei manager di produzione

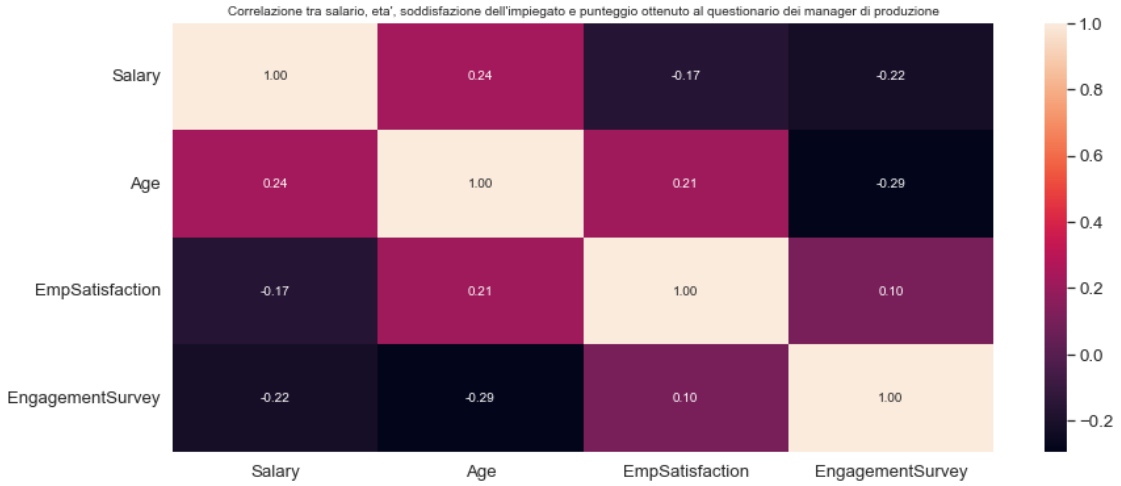
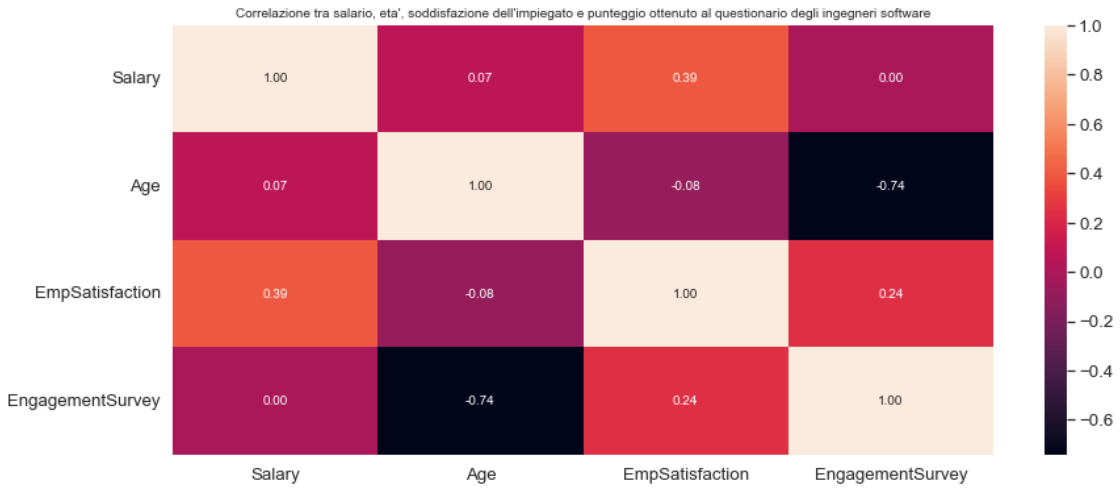


Grafico 15.5: Correlazione tra salario, eta', soddisfazione dell'impiegato e punteggio ottenuto al questionario degli ingegneri software



3.0.1 *Quartili e Outliers*

Di seguito creiamo, relativamente alle variiabili numeriche, dei dataset in cui sono stati rimossi gli outliers. Su questi dataset (confrontati con quelli aventi gli outliers) calcoliamo quartili e IQR.

[73]:
```
def remove_outlier(df_in, col_name):
    q1 = df_in[col_name].quantile(0.25)
    q3 = df_in[col_name].quantile(0.75)
    iqr = q3-q1
    fence_low  = q1-1.5*iqr
    fence_high = q3+1.5*iqr
    df_out = df_in.loc[(df_in[col_name] > fence_low) & (df_in[col_name] <␣
 ↪fence_high)]
    return df_out

def quartili_fence_iqr(df_in, col_name):
    q1 = df_in[col_name].quantile(0.25)
    q3 = df_in[col_name].quantile(0.75)
    iqr = q3-q1 #Interquartile range
    fence_low  = q1-1.5*iqr
    fence_high = q3+1.5*iqr
    d = {}
    d["q1"] = [q1]
    d["q3"] = [q3]
    d["iqr"] = [iqr]
    d["fence_low"] = [fence_low]
    d["fence_high"] = [fence_high]
    return d
```

Salario Differnza Interquartili, IQR

`Outlier compresi`

q1	q3	iqr	fence_low	fence_high
55501.5	72036.0	16534.5	30699.75	96837.75

`Senza Outliers`

q1	q3	iqr	fence_low	fence_high
54381.25	67221.75	12840.5	35120.5	86482.5

Age Differnza Interquartili, IQR

`Outlier compresi`

q1	q3	iqr	fence_low	fence_high
36.0	49.0	13.0	16.5	68.5

`Senza Outlier`

q1	q3	iqr	fence_low	fence_high
36.0	48.0	12.0	18.0	66.0

Age of Hire Differnza Interquartili, IQR

`Outlier compresi`

q1	q3	iqr	fence_low	fence_high
28.0	39.0	11.0	11.5	55.5

`Senza outlier`

q1	q3	iqr	fence_low	fence_high
36.0	48.0	12.0	18.0	66.0

Age of Term Differnza Interquartili, IQR

`Outlier compresi`

q1	q3	iqr	fence_low	fence_high
29.75	43.25	13.5	9.5	63.5

`Senza Outlier`

q1	q3	iqr	fence_low	fence_high
29.25	43.0	13.75	8.625	63.625

Assenze Differnza Interquartili, IQR

`Outlier compresi`

q1	q3	iqr	fence_low	fence_high
5.0	15.0	10.0	-10.0	30.0

`Senza outlier`

q1	q3	iqr	fence_low	fence_high
5.0	15.0	10.0	-10.0	30.0

Enggement Survey Differnza Interquartili, IQR

`Outlier compresi`

q1	q3	iqr	fence_low	fence_high
3.69	4.7	1.01	2.175	6.215

`Senza outlier`

q1	q3	iqr	fence_low	fence_high
3.735	4.7	0.965	2.2875	6.1475

Calcolo di indici Skweness e Kurtosis In questa sezione si analizza l'andamento delle distribuzioni. In particolare, testiamo la normalità della curva per specifiche variabili tenendo in considerazione il numero di outlier. Se il numero di outlier all'interno di una colonna dovesse superare il 15%, questi verranno rimossi per il calcolo. Per ogni variabile presa in esame viene mostrata la visualizzazione della distribuzione.

[92]:
```
#con il parametro mod viene specificata la modalità di return della fn
def count_outliers(df_in, colname, mod):
    q1 = df_in[colname].quantile(0.25)
    q3 = df_in[colname].quantile(0.75)
    iqr = q3-q1 #Interquartile range
    fence_low  = q1-1.5*iqr
    fence_high = q3+1.5*iqr
    new = df_in[(df_in[colname] < fence_low) | (df_in[colname] > fence_high)].
 ↪count()
    out_num = new[colname]
    if mod==0:
        return out_num
    if mod==1:
        return out_num*100/df_in[colname].count()
```

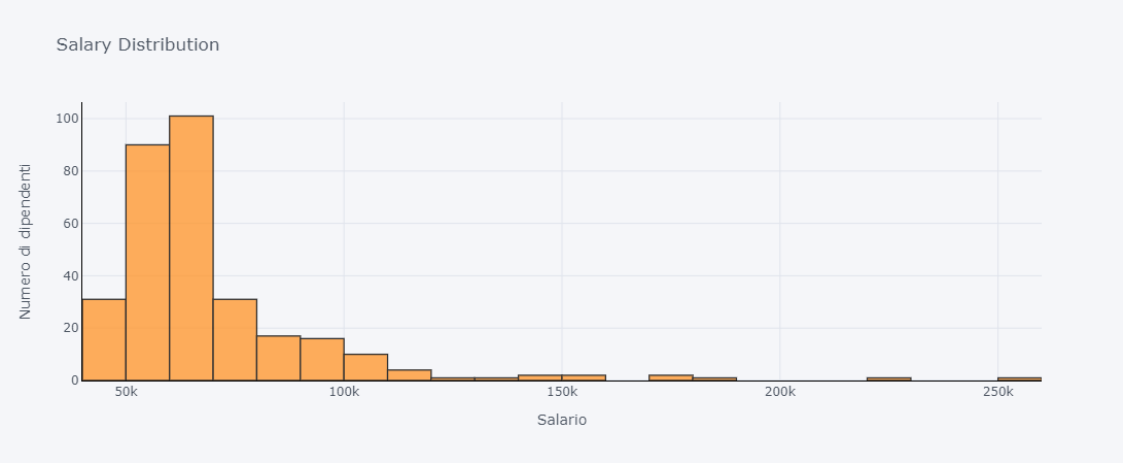
[93]:
```

```

```
#funzioni che calcolano kurtosis e skewness, tenendo in considerazione il numero␣
 ↪di outliers
def compute_kurt(df, colname, outliers):
    if outliers<=15:
        return round(df[colname].kurt(), 2)
    else:
        return round(remove_outlier(df, colname[colname]).kurt(), 2)

def compute_skew(df, colname, outliers):
    if outliers<=15:
        return round(df[colname].skew(), 2)
    else:
        return round(remove_outlier(df, colname[colname]).skew(), 2)

def kurt_skew_table(kurt, skew):
    d = {}
    d["Kurtosis"] = [kurt]
    d["Skewness"] = [skew]
    new_df = pd.DataFrame(data = d)
    return new_df
```

Salario

[94]:

	Kurtosis	Skewness
0	15.45	3.31



Age

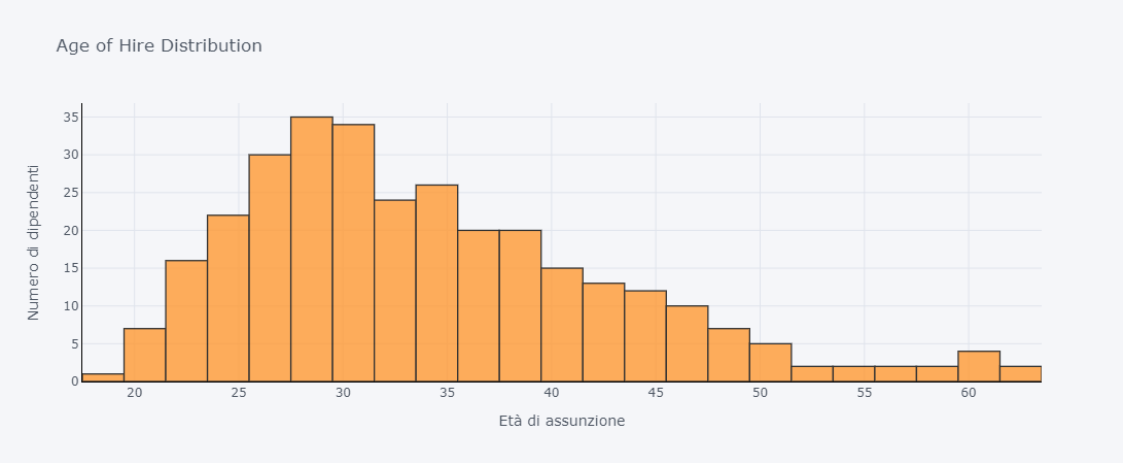
[96]:

	Kurtosis	Skewness
0	0.57	0.91

Age of Hire

[98]:

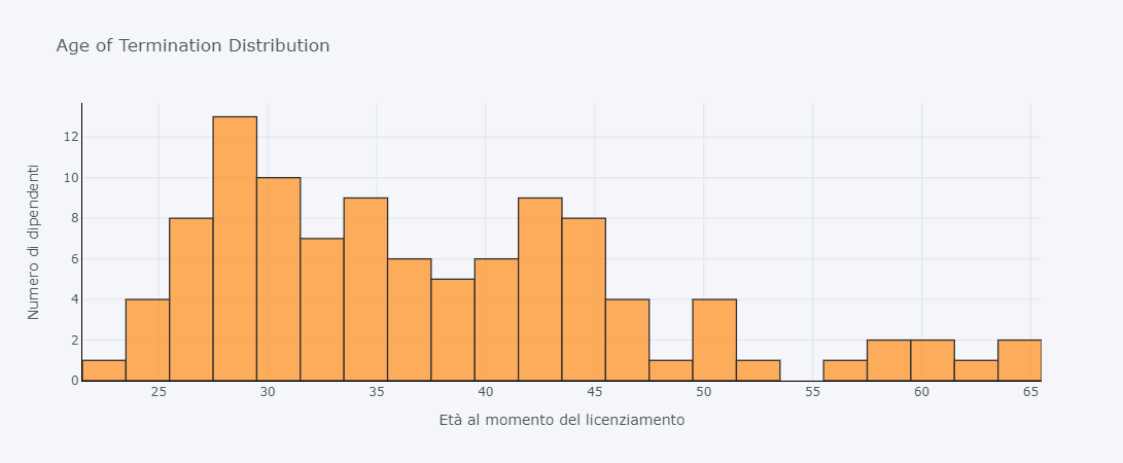
	Kurtosis	Skewness
0	0.54	0.89



Age of Term

[100]:

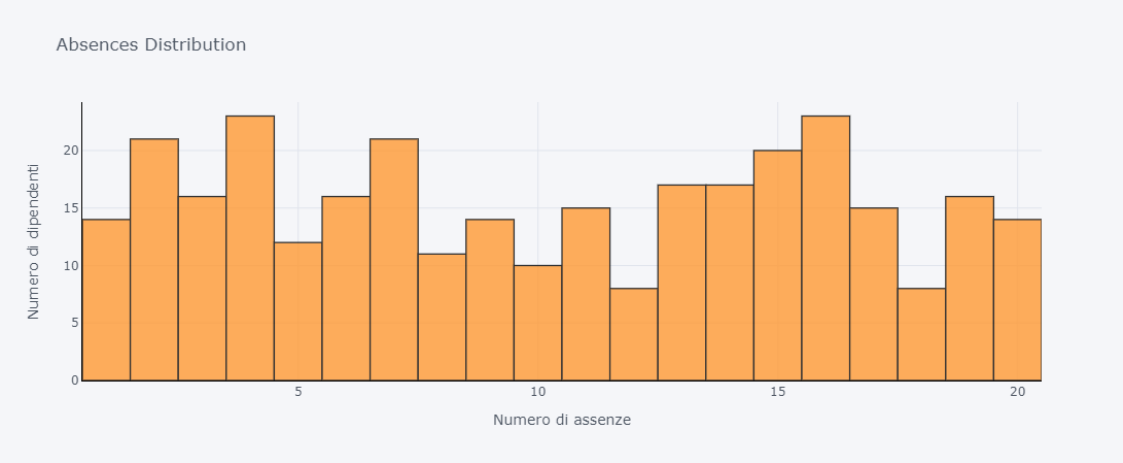
	Kurtosis	Skewness
0	0.28	0.87



Assenze

[102]:

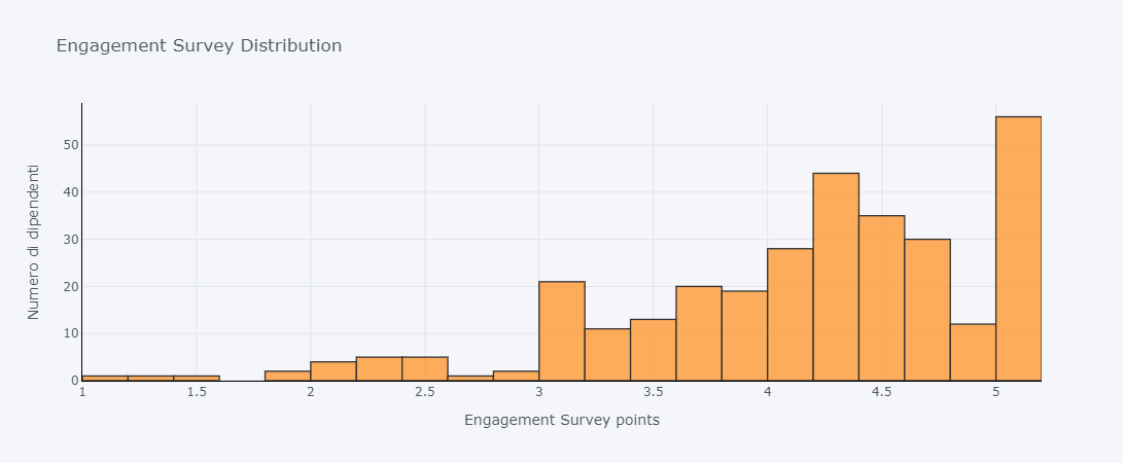
	Kurtosis	Skewness
0	-1.3	0.03



Engagement Survey

[104]:

	Kurtosis	Skewness
0	1.16	-1.12



Indici di correlazione di Pearson, Spearman, Kendall

Correlazione di Pearson

	Salary	Age	EmpSatisfaction	EngagementSurvey	\
Salary	1.000000	0.094360	0.062718	0.064966	
Age	0.094360	1.000000	-0.065241	0.063384	
EmpSatisfaction	0.062718	-0.065241	1.000000	0.187105	
EngagementSurvey	0.064966	0.063384	0.187105	1.000000	
Absences	0.082382	-0.039471	0.075222	-0.008771	
DaysLateLast30	-0.069443	-0.053286	-0.235412	-0.585232	
SpecialProjectsCount	0.508333	-0.090207	0.033877	0.013227	
AgeTerm	0.069028	0.980876	-0.033218	0.100420	
AgeofHire	0.124498	0.975796	-0.063188	0.057573	

	Absences	DaysLateLast30	SpecialProjectsCount	\
Salary	0.082382	-0.069443	0.508333	
Age	-0.039471	-0.053286	-0.090207	
EmpSatisfaction	0.075222	-0.235412	0.033877	
EngagementSurvey	-0.008771	-0.585232	0.013227	
Absences	1.000000	0.001833	-0.020452	
DaysLateLast30	0.001833	1.000000	-0.092494	
SpecialProjectsCount	-0.020452	-0.092494	1.000000	
AgeTerm	-0.082780	-0.148088	-0.040741	
AgeofHire	-0.054188	-0.058006	-0.005479	

	AgeTerm	AgeofHire
Salary	0.069028	0.124498
Age	0.980876	0.975796
EmpSatisfaction	-0.033218	-0.063188
EngagementSurvey	0.100420	0.057573

```
Absences              -0.082780  -0.054188
DaysLateLast30        -0.148088  -0.058006
SpecialProjectsCount  -0.040741  -0.005479
AgeTerm                1.000000   0.979593
AgeofHire              0.979593   1.000000
```

Correlazione di Spearman

```
                        Salary       Age  EmpSatisfaction  EngagementSurvey  \
Salary                1.000000  0.016838         0.040271          0.031117
Age                   0.016838  1.000000        -0.057629          0.044802
EmpSatisfaction       0.040271 -0.057629         1.000000          0.124273
EngagementSurvey      0.031117  0.044802         0.124273          1.000000
Absences              0.079194 -0.023349         0.073698         -0.006484
DaysLateLast30       -0.066890 -0.048690        -0.208183         -0.426366
SpecialProjectsCount  0.506866 -0.105184         0.011738          0.021082
AgeTerm               0.003070  0.969694        -0.018589          0.046344
AgeofHire             0.068585  0.963038        -0.058402          0.043293

                      Absences  DaysLateLast30  SpecialProjectsCount  \
Salary                0.079194       -0.066890              0.506866
Age                  -0.023349       -0.048690             -0.105184
EmpSatisfaction       0.073698       -0.208183              0.011738
EngagementSurvey     -0.006484       -0.426366              0.021082
Absences              1.000000       -0.003041             -0.022693
DaysLateLast30       -0.003041        1.000000             -0.072020
SpecialProjectsCount -0.022693       -0.072020              1.000000
AgeTerm              -0.026627       -0.090373             -0.066491
AgeofHire            -0.044665       -0.052332             -0.007273

                       AgeTerm  AgeofHire
Salary                0.003070   0.068585
Age                   0.969694   0.963038
EmpSatisfaction      -0.018589  -0.058402
EngagementSurvey      0.046344   0.043293
Absences             -0.026627  -0.044665
DaysLateLast30       -0.090373  -0.052332
SpecialProjectsCount -0.066491  -0.007273
AgeTerm               1.000000   0.966692
AgeofHire             0.966692   1.000000
```

Correlazione di Kendall

```
                        Salary       Age  EmpSatisfaction  EngagementSurvey  \
Salary                1.000000  0.011926         0.030117          0.021360
Age                   0.011926  1.000000        -0.045459          0.030435
EmpSatisfaction       0.030117 -0.045459         1.000000          0.098420
EngagementSurvey      0.021360  0.030435         0.098420          1.000000
```

```
Absences              0.053300 -0.014780         0.058468         -0.005388
DaysLateLast30       -0.053932 -0.039960        -0.189672         -0.349876
SpecialProjectsCount  0.393345 -0.083957         0.010048          0.016251
AgeTerm               0.001705  0.875579        -0.016384          0.028753
AgeofHire             0.045308  0.865595        -0.046153          0.030022

                      Absences  DaysLateLast30  SpecialProjectsCount  \
Salary                0.053300       -0.053932              0.393345
Age                  -0.014780       -0.039960             -0.083957
EmpSatisfaction       0.058468       -0.189672              0.010048
EngagementSurvey     -0.005388       -0.349876              0.016251
Absences              1.000000       -0.002432             -0.018979
DaysLateLast30       -0.002432        1.000000             -0.067538
SpecialProjectsCount -0.018979       -0.067538              1.000000
AgeTerm              -0.020003       -0.073248             -0.054092
AgeofHire            -0.031228       -0.042122             -0.005821

                       AgeTerm  AgeofHire
Salary                0.001705   0.045308
Age                   0.875579   0.865595
EmpSatisfaction      -0.016384  -0.046153
EngagementSurvey      0.028753   0.030022
Absences             -0.020003  -0.031228
DaysLateLast30       -0.073248  -0.042122
SpecialProjectsCount -0.054092  -0.005821
AgeTerm               1.000000   0.871424
AgeofHire             0.871424   1.000000
```

Considerando la quantità di outliers rilevati e la mancata presenza di distribuzioni normali (nonostante le distribuzioni di Age e AgeOfHire ci si avvicinino) all'interno del dataset, l'indice di correlazione più affidabile è quello di Kendall. Se osserviamo la tabella con le correlazioni, possiamo identificare, come ci si può facilmente aspettare, un'alta correlazione soltanto tra le variabili che indicano l'età e l'età di cessazione/età di assunzione; in quanto si tratta di dati derivati dalla medesima colonna. Per quanto riguarda le altre variabili non sono state rilevate correlazioni significative.

---

4 *Sezione Altro - Extra*

Questa ulteriore sezione mostra gli ultimi tre grafici che, nonostante non fossero direttamente legati alle principali domande poste in introduzione , ci sono sembrati interessanti da mostrare nella relazione. In primis ci siamo chiesti se lo stato civile potesse in qualche modo influenzare il numero delle assenze dal luogo di lavoro. Nel grafico 17, sempre relativamente alle assenze, abbiamo analizzato se vi fosse un rapporto tra queste e il sesso dell'impiegato. Questo avrebbe potuto aprire ad un'interessante analisi di questo dato rispetto allo stato di genitorialità degli impiegati; purtroppo però nel dataset non sono presenti queste informazioni. Infine, abbiamo cercato di capire se vi fosse una relazione tra la motivazione della cessazione del contratto e l'età

dell'impiegato.

Grafico 16: Rapporto tra assenze e stato civile Nel successivo barplot è stato evidenziato il numero di *assenze medio* in base allo stato civile. Nonostante non si presenti un'eccessiva differenza nel numero di assenze è interessante evidenziare come gli impiegati facenti parte della categoria *separati* abbiano in media accumulato meno assenze rispetto agli altri.

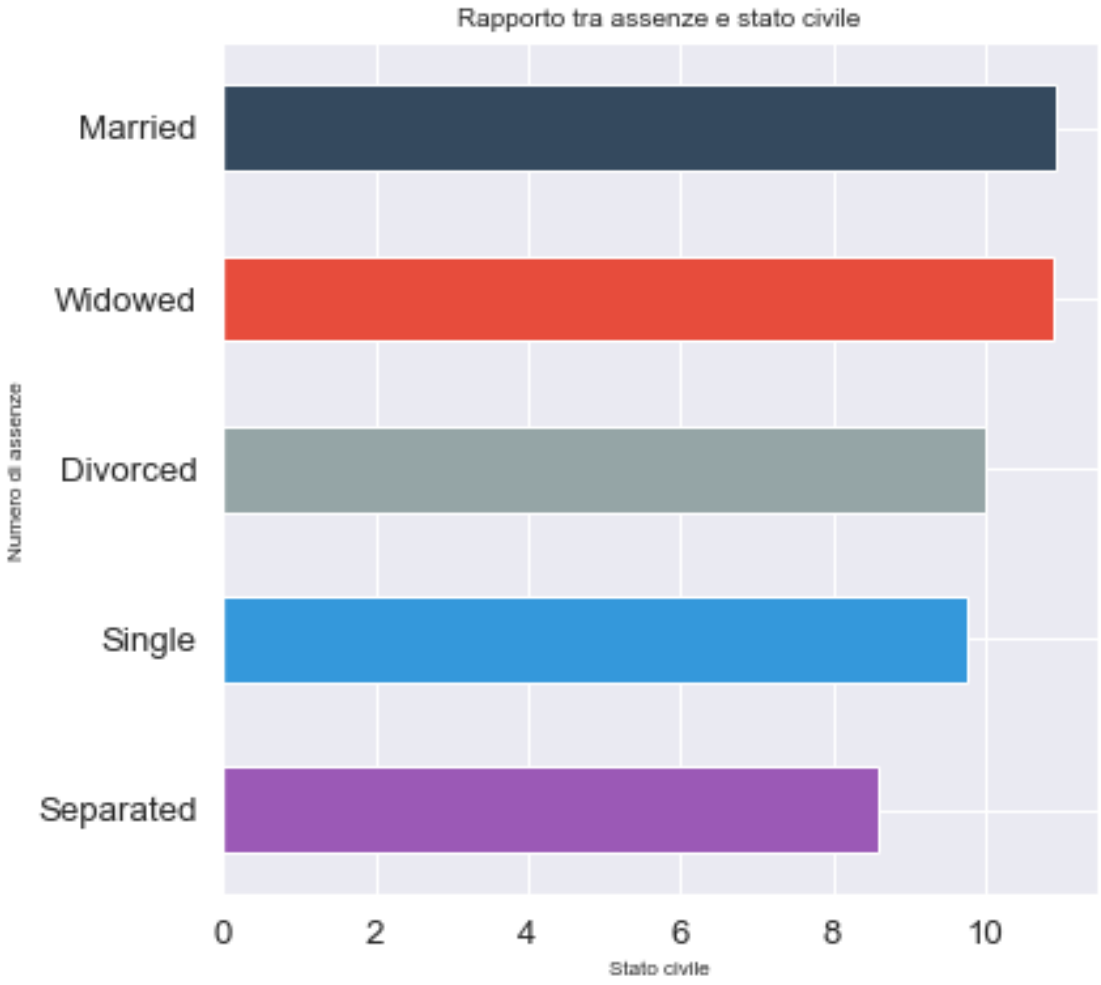


Grafico 17: Numero di assenze specifiche per uomini e donne, per ogni dipartimento In questo grafico effettuiamo un confronto tra le assenze di ogni impiegato e il dipartimento lavorativo. Nella maggior parte dei dipartimenti gli uomini fanno meno assenze delle donne, ad eccezione di quelli del dipartimento *Admin Offices* che hanno una media di quasi 11 assenze rispetto alle circa 8 delle donne. Anche in questo grafico, come nel grafico 10, si conferma l'assenza di uomini nella categoria "Executive Office". La CEO dell'azienda ha una media di 10 assenze.

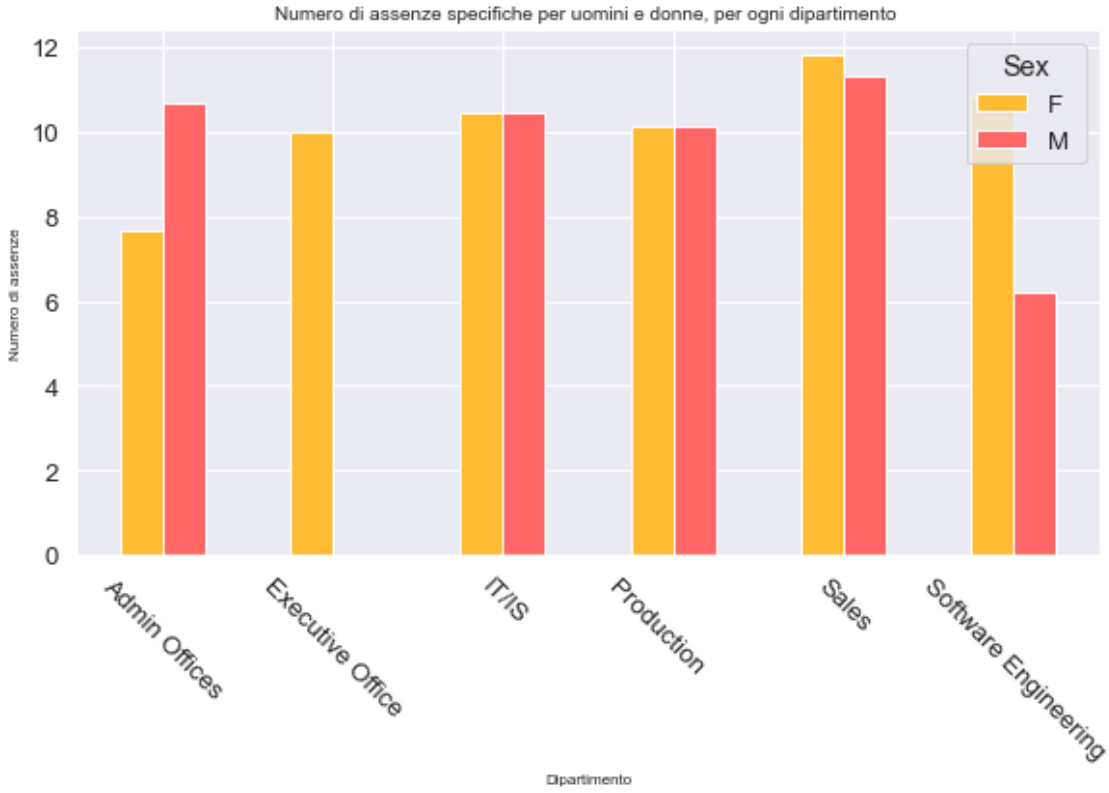
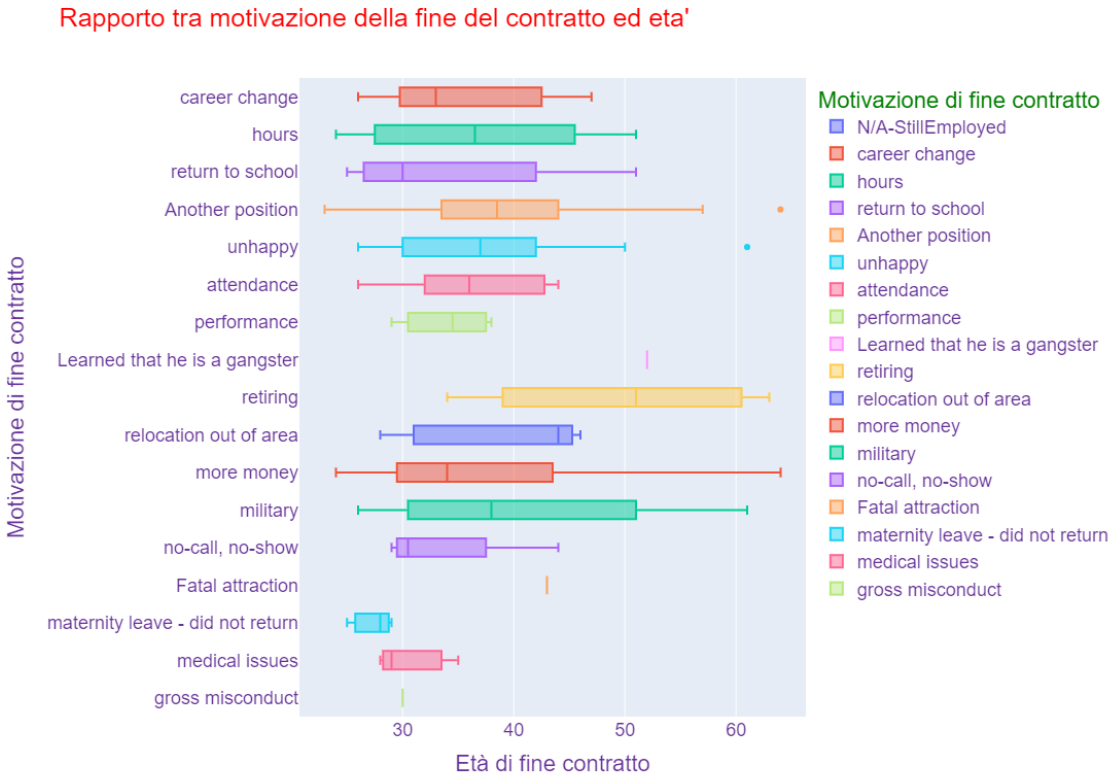


Grafico 18: Rapporto tra motivazione della fine del contratto ed eta' Nel seguente grafico possiamo osservare dei boxplot che rappresentano le distribuzioni della variabile "età" organizzati in base ai diversi motivi di cessazione del contratto di lavoro. Tra le distribuzioni più interessanti possiamo osservare il box plot creato per la motivazione *maternity leave*, in cui notiamo che i valori si concentrano in un'età molto giovane che non va oltre i trenta.

Motivazione di fine contratto	Numero di impiegati
N/A-StillEmployed	207
Another position	20
unhappy	14
more money	11
career change	9
hours	8
attendance	7
return to school	5
relocation out of area	5
no-call, no-show	4
military	4
retiring	4
performance	4
maternity leave - did not return	3

Motivazione di fine contratto	Numero di impiegati
medical issues	3
Learned that he is a gangster	1
Fatal attraction	1
gross misconduct	1



5 *Sezione Gestione dei NaN*

Visto e considerato che gli unici dati non numerici erano presenti nella colonna "DateofTermination", abbiamo sostuito queste occorrenze con "Still Employed", poichè così come riportato su Kaggle i dati NaN erano riferiti a quegli impiegati ancora in esercizio

[112]:
```
#df.DateofTermination = df.DateofTermination.fillna('Still_Employed',␣
↪inplace=True)
df.DateofTermination = df.DateofTermination.fillna('Still_Employed')
df.AgeTerm = df.AgeTerm.fillna('Still_Employed')
```

6 *Conclusioni*

L'analisi di questo dataset ci ha permesso di giungere alle seguenti conclusioni:

- il salario dei dipendenti corrisponde con la posizione nell'organigramma aziendale
- le persone con gli stipendi più alti non eseguono un gran numero di progetti speciali
- vi un forte rendimento presente nell'azienda nonostante il numero di assenze che ammonta a un minimo di 10 ore
- la maggioranza dei dipendenti sono donne e il CEO è una donna
- la maggior parte dei dipendenti hanno situazioni familiari stabili e sono impegnati
- quasi tutti i dipendenti sono stati assunti attraverso piattaforme online (Indeed, LinkedIn, etc.)
- non vi è una forte relazione tra la piattaforma di assunzione e il punteggio di rendimento del dipendente
- a fronte del calcolo degli indici di correlazione si può affermare che non vi sono correlazioni significative (in quanto nessuno dei valori presenti nelle tabelle sopraripotate si avvicina abbastanza a +1 o -1)
- riguardo i dati quantitativi analizzati si può altresì affermare che non vi sono distribuzioni normali dei valori

6.1 *ChangeLog*

Negli elenchi seguenti si riportano modifiche e aggiunte rispetto alla precedente versione del progetto

Modifiche

- grafico 3 (stile)
- grafico 4 (stile)
- grafico 8 (trasformazione in grafico interattivo)
- grafico 9 (trasformazione in grafico interattivo)
- grafico 10 (trasformazione in grafico interattivo)
- grafico 12->13 (trasformazione in grafico interattivo)
- grafico 13->14 (trasformazione in grafico interattivo)
- grafici 14.1->15.1, 14.2->15.2, 14.3->15.3, 14.4->15.4, 14.5->15.5 (cambio dell'indice di correlazione nelle heatmap: Pearson->Kendall)
- grafico 15->16 (stile)
- grafico 17->18 (trasformazione in grafico interattivo)

Aggiunte

- sezione con statistiche descrittive sulle variabili numeriche del dataset
- grafico 12: bubbleplot sul rapporto tra salario, assenze e progetti speciali in base all'età
- grafico 15: heatmap su variabili numeriche
- sottosezione *quartili e outliers*
- sottosezione *kurtosis e skewness* con visualizzazione delle distribuzioni
- sottosezione con i tre indici di correlazione per il rapporto tra ciascuna variabile numerica nel dataset
- sezione *gestione dei NaN*
- aggiunta conclusioni