Università degli Studi di Milano-Bicocca

Anno accademico 2020-2021

CdLM in Data Science

Esame di Foundations of Probability and Statistics

Andrea Afify (Matricola: 813466)

Marta La Franca (Matricola: 866590)

Alessandro Risaro (Matricola:825113)

Indice

Abstract

In questo lavoro si analizza il dataset relativo alla Rilevazione sulle Forze di Lavoro condotto dall'Istat con l'obiettivo da una parte di stimare i livelli di occupazione, disoccupazione, attività e inattività nel Paese, dall'altra quello di stimare la retribuzione media in funzione di alcune caratteristiche demografiche come ad esempio area geografica, livello di istruzione e genere dei rispondenti alla rilevazione.

Innanzitutto dopo aver esplorato il dataset viene effettuata un'analisi descrittiva univariata delle sopra citate variabili demografiche. In seguito si effettua un'analisi bivariata da una parte relativa al tasso di occupazione, disoccupazione, attività e inattività in relazione alle variabili demografiche dall'altra della retribuzione dei rispondenti in funzione di genere,titolo di studio, posizione lavorativa ed età. Per verificare la potenziale associazione tra le variabili si effettua il test del chi quadro.

Infine viene condotta una parte di analisi inferenziale. In particolare, si effettua l'analisi della varianza per la variabile relativa alla retribuzione mensile in relazione ad alcune variabili categoriche di interesse. Si utilizza poi il modello di regressione multipla per vedere l'influenza delle variabili demografiche sulla retribuzione dei rispondenti. Viene ricercata un'eventuale presenza di multicollinearità utilizzando il fattore di inflazione della varianza (VIF).

1. Introduzione

Dal gennaio 2004 l'Istat pubblica, ogni trimestre, una *Rilevazione sulle Forze di Lavoro (RFL*).

Si tratta di una rilevazione campionaria svolta con la finalità di ottenere informazioni statistiche sul mercato del lavoro in Italia ed in particolare: sulla situazione lavorativa, sulla ricerca di lavoro e sugli atteggiamenti verso il mercato del lavoro della popolazione in età lavorativa.

I risultati dell'indagine vengono diffusi attraverso comunicati stampa e tavole di dati e sono disponibili dal primo trimestre del 2014. L'Istat rende disponibili i file mIcro.STAT della Rilevazione sulle Forze di Lavoro al link https://www.istat.it/it/archivio/127792.

In questo elaborato vengono analizzati i microdati provenienti dalla Rilevazione sulle Forze di lavoro del *secondo trimestre del 2020 pubblicati il 2 ottobre 2020*.

L'analisi di tali dati si concentrata in tre fasi:

- *analisi descrittiva univariata e bivariata di alcune variabili individuate;*

- *calcolo dei principali indicatori relativi al lavoro*: tasso di occupazione, tasso di attività, tasso di disoccupazione, tasso di inattività valutati sia singolarmente che in relazione alla variabile genere;

- *analisi della retribuzione mensile dei lavoratori occupati dipendenti, in relazione ad alcune variabili,* attraverso: analisi descrittiva univariata e bivariata, test del chi-quadrato, analisi della varianza (ANOVA), VIF e costruzione di un modello di regressione multipla.

2. Informazioni sulla rilevazione: il questionario

I dati della *Rilevazione sulle Forze di Lavoro* provengono da un'indagine campionaria attraverso la quale, ogni anno, vengono intervistate *250 mila* famiglie residenti in Italia distribuite in circa *1.400* comuni italiani. I dati vengono ottenuti attraverso la somministrazione di un questionario a tutti i componenti (non minori di 15 anni di età) delle famiglie. Il questionario è diviso in *sezioni* che sono di seguito elencate:

- *sezione SG (Scheda Generale)*: vengono rilevate informazioni anagrafiche e informazioni riguardanti il titolo di studio dell'intervistato (per tutti i componenti della famiglia);

- *sezione A*: vengono rilevate informazioni anagrafiche di chi risponde al questionario (componenti della famiglia che hanno 15 anni o più):

- *sezione B:* situazione lavorativa nella settimana di riferimento (per le persone con 15 anni o più);

- *sezione C*: solo con riferimento alle persone occupate, vengono rilevate informazioni relative alle caratteristiche del rapporto di lavoro, del l'attività economica dell'impresa in cui essi lavoro, la professione svolta, la tipologia di orario svolto e altre informazioni rilevanti;

- *sezione D*: solo con riferimento alle persone occupate, vengono rilevate informazioni relative ad attività lavorative secondarie;

- *sezione E*: solo con riferimento alle persone non occupate, vengono rilevate informazioni relative alle precedenti esperienze lavorative;

- *sezione F*: solo con riferimento alle persone di 15 anni o più occupati o non occupati, vengono rilevate informazioni relative alla ricerca del lavoro;

- *sezione G*: solo con riferimento alle persone con un'età tra i 15 e i 74, vengono rilevate informazioni relative all'utilizzo di servizi per l'impiego e agenzie per il lavoro;

- *sezione H*: solo con riferimento alle persone di 15 anni o più, vengono rilevate informazioni relative all'istruzione e formazione professionale;

- *sezione I*: solo con riferimento alle persone con 15 anni o più, vengono rilevate informazioni relative alla condizione autopercepita relativa all'anno in corso e all'anno precedente oltre all'informazione in merito alla residenza.

3. Esplorazione del data set e analisi descrittiva

Il dataset fornito dall' Istat è composto da *101.600* record (corrispondenti al numero di persone che hanno risposto al questionario) e *343* variabili. Per l'analisi sono state selezionate solo alcune variabili rilevanti, relative *perlopiù* alle sezioni *SG, B e C*:

3.1: sezione SG: Informazioni anagrafiche e riguardanti il titolo di studio

- *variabile RIP5*: ripartizione dei rispondenti in cinque aree geografiche:

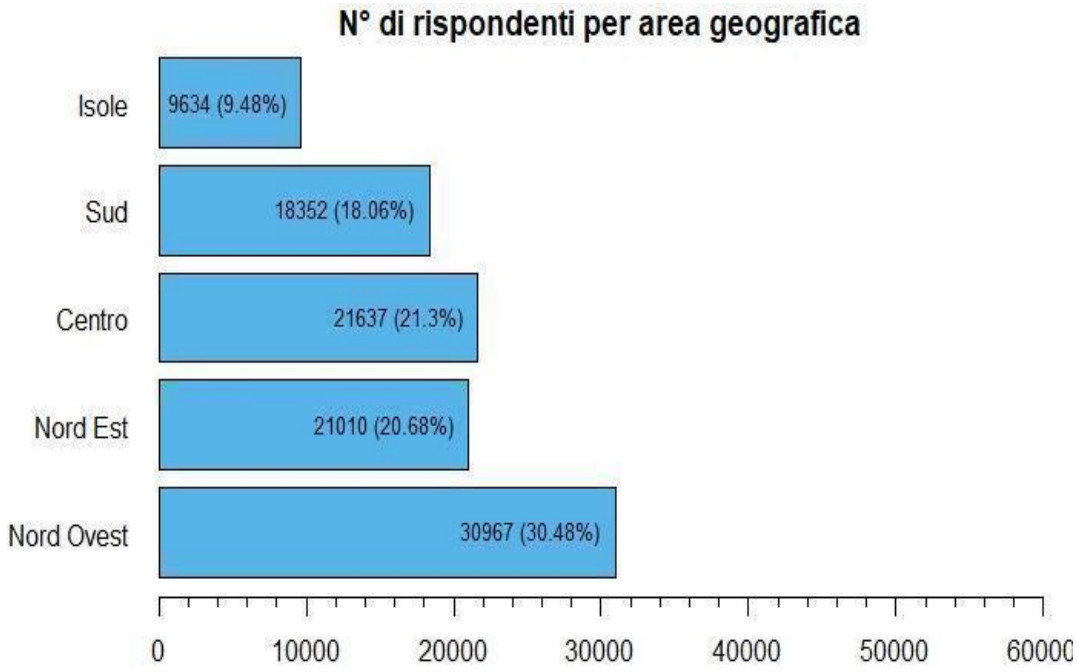

Figura.1: In questa figura si mostra il numero di rispondenti per area geografica. Per maggiore chiarezza si è scelto di distinguere le aree secondo le cinque modalità della variabile RIP5.

- *variabili TN2 e RPN2*: rispettivamente tipologia di nucleo familiare e relazione di parentela all'interno della famiglia

TN2	Frequenze assolute	%
Persona isolata	17949	17.67
Coppia con figli	48031	47.27
Coppia senza figli	26760	26.34
Monogenitore maschio	1357	1.34
Monogenitore femmina	7503	7.38

- Tabella 1

RPN2	Frequenze assolute	%
Persona singola	17949	17.67
Capo nucleo	30508	30.03
Coniuge o convivente	26686	26.27
Figlio	26454	26.04

- Tabella 2

- *variabile SG4*: numero di persone che vivono in casa:

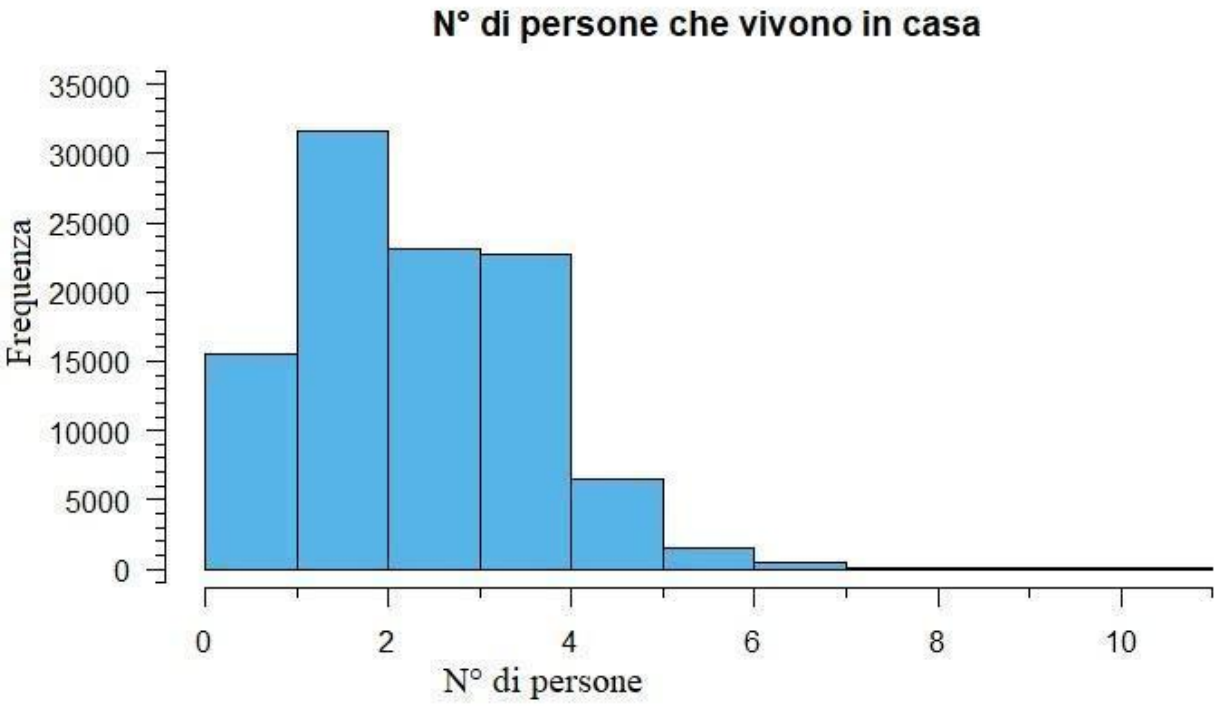

Figura.2: In questa figura si mostra l'istogramma della variabile SG4.

- *variabile SG11*: genere dei rispondenti

SG11	%
Maschi	47.46
Femmine	52.54

- Tabella 3

- *variabile ETAM*: età dei rispondenti

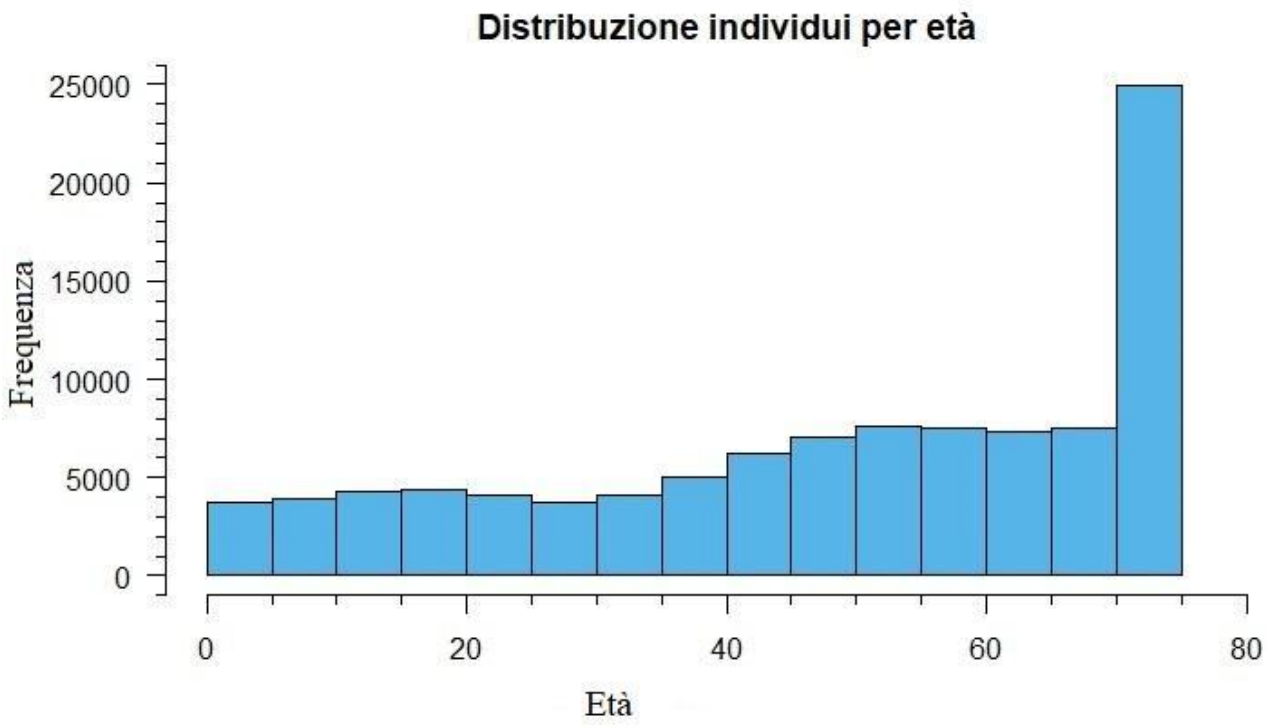

Figura.3: In questa figura si mostra l'istogramma della variabile ETAM.

- *variabile CLETAD*: rispondenti per classe di età:

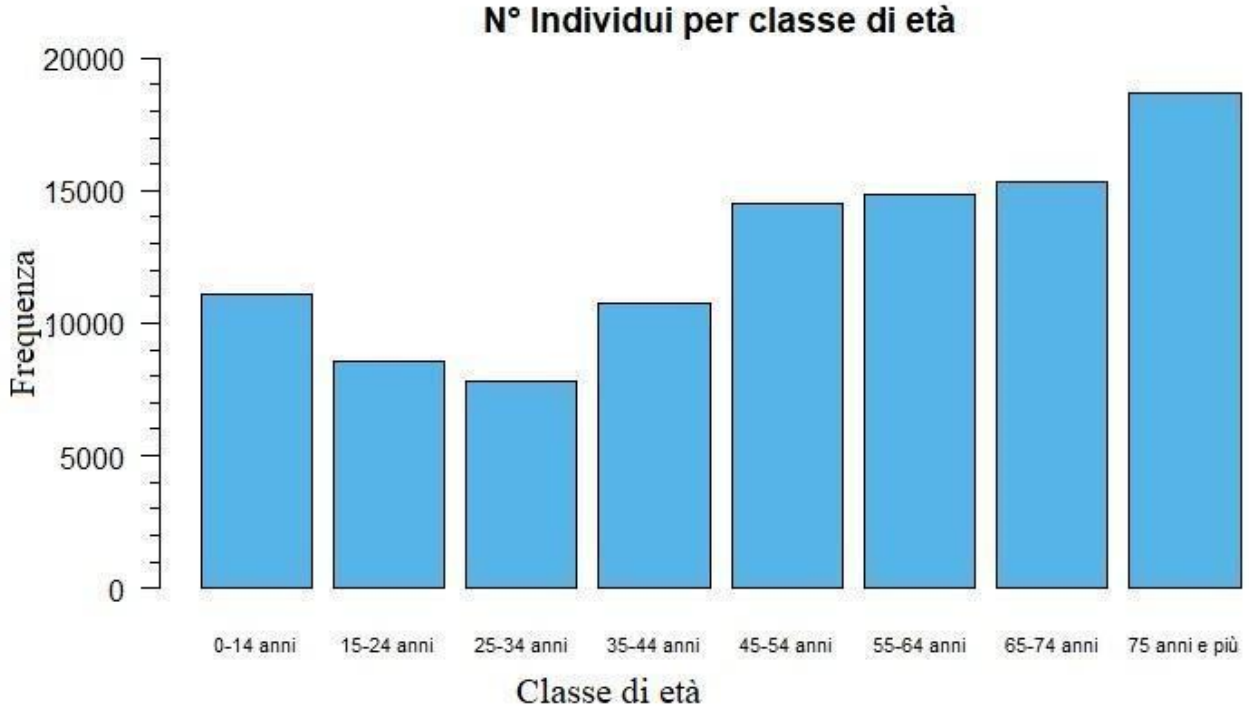

Figura.4: In questa figura si mostra il barplot della variabile CLETAD.

- *variabile STACIM*: stato civile dei rispondenti

STACIM	Frequenze assolute	%
Celibe/nubile	36273	35.70
Coniugato/a	48918	48.15
Separato/a o divorziato/a	6478	6.38
Vedevo/a	9931	9.77

-Tabella 4

- *variabile SG13*: luogo di nascita dei rispondenti

SG13	%
Italia	91.76
Estero	8.24

- Tabella 5

- *variabile CITTAD*: Cittadinanza dei rispondenti

CITTAD	Frequenze assolute	%
Italiana	94663	93.17
UE	2190	2.16
Fuori UE	4747	4.67

- Tabella 6

- *variabili SG18B e SG18F*: rispettivamente anno in cui il rispondente è venuto a vivere in Italia e anno dal quale vive in Italia senza allontanarsi

SG18B	Min	Q1	Mediana	Media	Q3	Max
	1928	1996	2003	1999	2008	2020

- Tabella 7

SG18F	Min	Q1	Mediana	Media	Q3	Max
	1953	1997	2007	2003	2014	2020

- Tabella 8

- *variabile EDULEV*: titolo di studio dei rispondenti in 6 modalità:

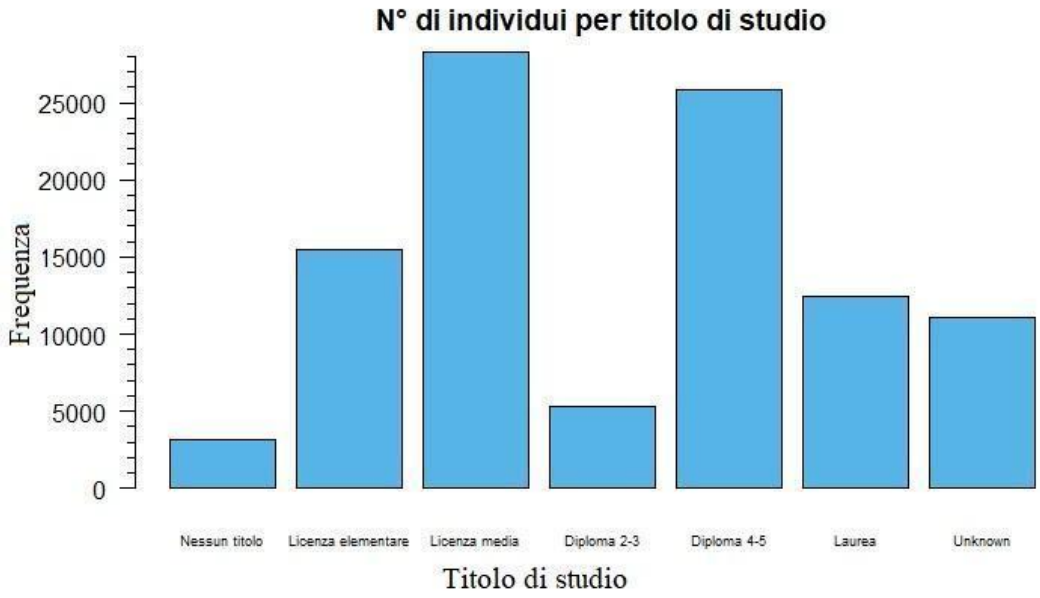

Figura.5: In questa figura si mostra il barplot della variabile EDULEV.

- *variabili SG24A, SG24B, SG26*: rispettivamente titolo di diploma conseguito dai rispondenti, titolo di studio post-laurea conseguito dai rispondenti, anno di conseguimento del titolo (sono esclusi dall'analisi i valori NA)

SG24A	%
Diploma vecchio ordinamento	65.15
Diploma AFAM primo livello	18.94
Diploma AFAM secondo livello	15.53

- Tabella 9

SG24B	%
Master di primo livello	3.78
Master di secondo livelo	1.29
Diploma di specializzazione univeristaria	4.55
Dottorato di ricerca	2.09
Nessuno di questi	88.30

- Tabella 10

SG26	Min	Q1	Mediana	Media	Q3	Max
	1927	1966	1983	1984	2002	2020

- Tabella 11

- *variabili COND3, COND10*: situazione professionale dei rispondenti rispettivamente a 3 modalità e a 10 modalità

COND3	Frequenze assolute	%
Occupati	34028	33.49
Persone in cerca di lavoro	2474	2.44
Inattivi	65098	64.07

- Tabella 12

COND10	Frequenze assolute	%
Occupati	34028	33.49
Persone ex-occupate in cerca, con precendenti esperienze	1275	1.25
Persone ex-inattivi, in cerca con precedenti esperienze	652	0.64
Persone in cerca, senza precendenti esperieze	547	0.54
Inattivi in età lavorativa, cercano non attivamente ma disponibili	2564	2.52
Inattivi in età lavorativa, cercano ma non disponibili	559	0.55
Inattivi in età lavorativa, non cercano ma disponibili	1954	1.92
Inattivi in età lavorativa, non cercano e non disponibili	16285	16.03
Inattivi in età non lavorativa, meno di 15 anni	11072	10.90
Inattivi in età non lavorativa, più di 64 anni	32664	32.15

- Tabella 13

3.2: sezione B: situazione lavorativa nella settimana di riferimento

- *variabile B1*: svolgimento di almeno un'ora di lavoro nella settimana di riferimento

B1	Frequenze assolute	%
Si	27007	26.58
No	62177	61.20
Permanentemente inabile al lavoro	1344	1.32
NA	11072	10.90

-Tabella 14

- *variabile B2*: ha un lavoro, ma non l'ha svolto nella settimana di riferimento

B2	Frequenze assolute	%
Si	7231	7.12
No	54946	54.08
NA	39423	38.30

- Tabella 15

- *variabile B3*: motivo per cui non si ha lavorato (sono esclusi dall'analisi i valori NA)

B3	%
Cassa integrazione	35.36
Ridotta attività dell'impresa	12.22
Controversia di lavoro	0.01
Maltempo	0.03
Malattia o infortunio	2.55
Ferie	6.49
Festività nella settimana	3.27
Orario variabiile o flessibile	0.07
Part-time verticale	0.01
Studio non riconosciuto	2.04
Studio non riconosciuto nell'orario lavorativo	0.91
Maternità obblligatoria	0.91
Maternità facoltativa	2.70
Motivi familiari	0.04
Mancanza/scarsità di lavoro	0.08
Lavoro occasionale	0.29
Lavoro stagionale alle dipendenze	32.82

- *Tabella 16*

3.3: sezione C: attività lavorativa principale (per gli occupati)

- *variabile C1*: tipologia di lavoro (sono esclusi dall'analisi i valori NA)

C1	%
Lavoro dipendente	75.80
Collaborazione coordinata e continuativa	0.48
Prestazione d'opera occasionale	0.28
Imprenditore	1.19
Libero professionista	6.24
Lavoratore in proprio	14.02
Coadiuvante in azienda familiare	1.57
Socio di cooperativa	0.42

- *Tabella 17*

- *variabile C9*: livello professionale (sono esclusi dall'analisi i valori NA)

C9	%
Dirigente	2.14
Quadro	6.78
Impiegato	45.11
Operaio	45.25
Apprendista	0.70
Lavoratore presso proprio domicilio per conto di un'impresa	0.03

- *Tabella 18*

- *variabile CAT12*: attività economica 12 classi (sono esclusi dall'analisi i valori NA)

CAT12	%
Agricoltura, sivicoltura, pesca	4.02
Industria in senso stretto	20.46
Costruzioni	5.95
Commercio	13.51
Alberghi e ristoranti	5.61
Trasporto e immagazzinaggio	4.72
Servizio di informazione e comunicazione	2.25
Attività finanziarie e assicurative	2.68
Attività immobiliari, servizi delle imprese	11.10
Amministrazione pubblica e difesa	5.82
Istruzione, sanità e altri servizi sociali	16.88
Altri servizi collettivi e personali	7.01

- *Tabella 19*

- *variabile C18*: numero di lavoratori della propria sede (sono esclusi dall'analisi i valori NA)

C18	%
Fino a 10 persone	32.03
Da 11 a 15 persone	9.31
Da 16 a 19 persone	3.73
Da 20 a 49 persone	14.31
Da 50 a 249 persone	20.30
250 persone o più	11.65
Non sa ma fino a 10 persone	2.63
Non sa ma più di 10 persone	6.05

- *Tabella 20*

- *variabile C20*: tipologia di contratto (sono esclusi dall'analisi i valori NA)

C20	%
Contratto a tempo determinato	13.28
Contratto a tempo indeterminato	86.72

- Tabella 21

- *variabile DURATT*: durata del lavoro attuale in mesi

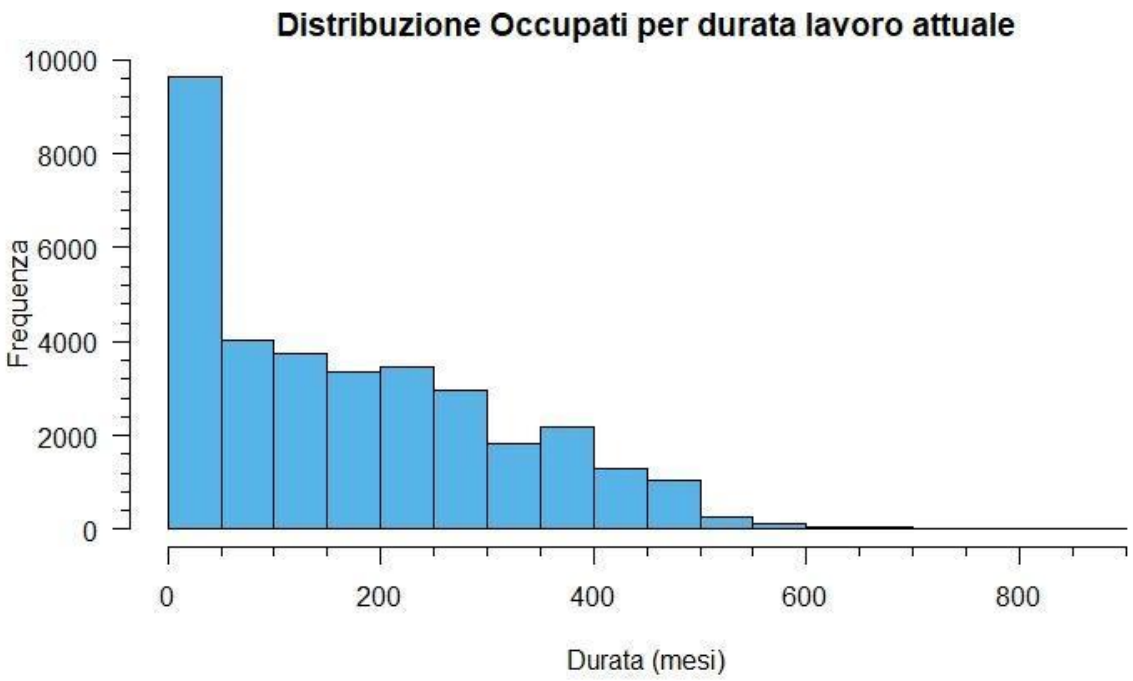

Figura.6: In questa figura si mostra l'istogramma della variabile DURATT.

- *variabile C27*: lavoro a tempo pieno o part-time

C27	%
Lavoro a tempo pieno	81.56
Part-time	18.44

- Tabella 22

- *variabile ORELAV*: Numero di lavoro effettivo a settimana

ORELAV	Min	Q1	Mediana	Media	Q3	Max	Dev.Std. Campionaria
	0.00	12.00	36.00	32.11	40.00	105.00	1.784.088

- Tabella 23

- *variabile RETRIC*: retribuzione netta del mese precedente

RETRIC	Mediana	Media	Dev.Std. Campionaria
	1330	1351	536.86

- Tabella 24

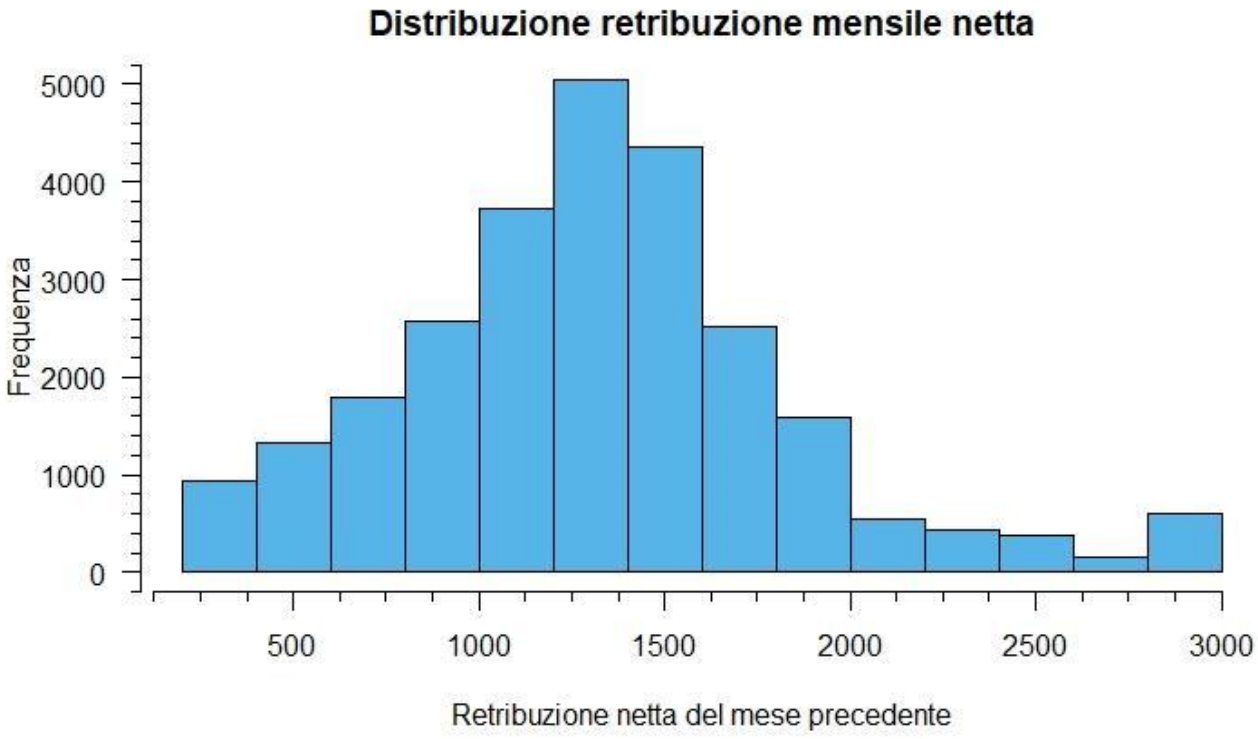

Figura.7: In questa figura si mostra l'istogramma della variabile RETRIC.

- *livello di soddisfazione*: grado di soddisfazione del lavoro rispetto diverse variabili

Variabile	Min	Q1	Mediana	Media	Q3	Max
Soddisfazione lavoro attuale	0	7	8	7.54	9	10
Soddisfazione guadagno attuale	0	6	7	6.69	8	10
Soddisfazione ambiente lavorativo	0	7	8	7.66	9	10
Soddisfazione carriera	0	5	7	6.27	8	10
Soddisfazione ore di lavoro	0	6	8	7.21	8	10
Grado di stabilità del lavoro	0	6	8	7.45	9	10
Grado di interesse verso il lavoro	0	7	8	8.07	10	10
Grado di facilità del lavoro	1	2	2	1.89	2	2

- Tabella 25

3.4: sezione D: informazioni relative ad attività secondarie (solo per occupati)

- *variabile D1*: secondo lavoro

D1	%
Si, uno	1.40
Si, più di uno	0.03
No	98.57

- Tabella 26

4. Tasso di occupazione

Valutando gli occupati *(33.49%)* in cerca di lavoro (*2.44%)* e inattivi *(64.07%)* - *variabile COND3*- e, più in particolare, gli individui in età lavorativa *(i.e. le persone tra i 15 e i 64 anni)*,è possibile ricavare importanti dati riguardanti l'analisi occupazionale in Italia ovvero:

- *tasso di attività*: rapporto tra popolazione attiva e popolazione in età lavorativa;
- *tasso di inattività*: rapporto tra popolazione attiva e popolazione in età lavorativa;
- *tasso di occupazione*: rapporto percentuale tra il numero di persone occupate e la popolazione;
- *tasso di disoccupazione*: rapporto percentuale tra il numero dei disoccupati e il totale della forza lavoro.

Esaminando i valori corrispondenti alla voci sopra elencate, viene determinata una stima intervallare con un test a due code e con un grado di fiducia del 95%, considerando una distribuzione binomiale *B(n,p)* di parametri *p* (oggetto di studio) e *n* (numerosità campione).

Indice	Limite inferiore	Limite superiore
Tasso di attività	64.08%	64.90%
Tasso di occupazione	57.62%	58.46%
Tasso di inattività	35.10%	35.91%
Tasso di disoccupazione	9.70%	10.33%

- *Tabella 27*

Viene inoltre calcolato l'intervallo di confidenza della differenza tra il tasso di attività e il tasso di occupazione maschile e femminile:

Indice	Gruppo	Limite inferiore	Limite superiore
Tasso di attività	Maschi/Femmine	11.28%	12.46%
Tasso di occupazione	Maschi/Femmine	11.02%	12.18%

- *Tabella 28*

Le differenze superano il *10%*, evidenziando la presenza di una disparità di genere nel contesto lavorativo italiano.

5. Analisi della retribuzione mensile

Nel seguente capitolo viene approfondita l'analisi della *variabile RETRIC*, variabile quantitativa continua, che rappresenta la retribuzione mensile dei lavoratori dipendenti.

Ai fini di tale analisi vengono considerati esclusivamente:

- *I lavoratori dipendenti*
- *Le persone in età lavorativa (dai 15 ai 64 anni)*

Si tratta di un totale di *25792* individui ed assume valori compresi in un range [0,3000)

5.1. Analisi univariata

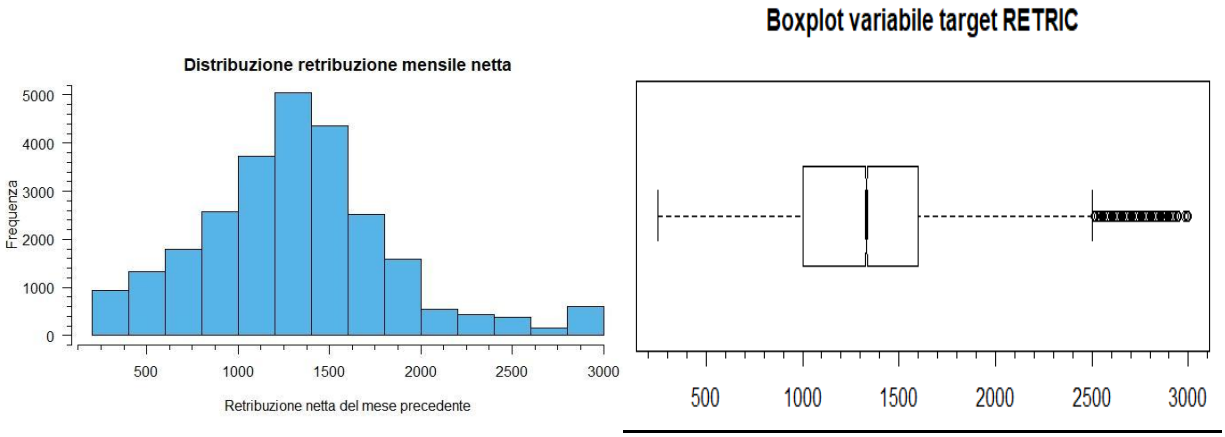

Figura.8: In questa figura si mostra nel panel di sinistra l'istogramma della variabile RETRIC, nel panel di destra il box-plot della stessa variabile.

E’ immediato osservare come, all’interno del range, la distribuzione della variabile target presenti una forma a campana, simile a quella di una normale.

Tuttavia, la maggiore distribuzione in corrispondenza del valore 3.000, è frutto di un’approssimazione all’interno del dataset per cui, tutti i valori superiori a 3.000 sono riportati come pari a 3.000. Utilizzando i valori corretti delle retribuzioni, la coda sarebbe quindi più lunga e la media della distribuzione sarebbe maggiore.

RETRIC	Min	Q1	Mediana	Media	Q3	Max	Dev. Std. Campionaria	Moda
	250	1000	1330	1352	1600	3000	537.2	1500

- Tabella 29

Dall’analisi degli indici di distribuzione si osserva che, la differenza del valore della media e della mediana è trascurabile (bisogna però ricordare che la media è sottostimata a causa dell’approssimazione dei valori superiori a 3.000), mentre il valor della moda risulta superiore ad entrambi: la variabile presenta, infatti, una leggera asimmetria negativa, che è possibile osservare anche considerando il box-plot nella figura a destra.

Ciononostante, nel seguito dell’analisi, si considera comunque valida l’ipotesi di approssimazione con una normale.

5.2 Analisi bivariata

Viene ora effettuata l’analisi della *variabile RETRIC* considerandola rispetto a diverse variabili categoriche del dataset e, attraverso un’analisi dei box-plot per evidenziare le eventuali dipendenze.

- *analisi variabile RETRIC rispetto alla variabile SG11 (genere)*

SG11	Mediana	Media	Dev. Std. Campionaria
Maschio	1400	1476.33	533.02
Femmina	1200	1214.16	507.3

- Tabella 30

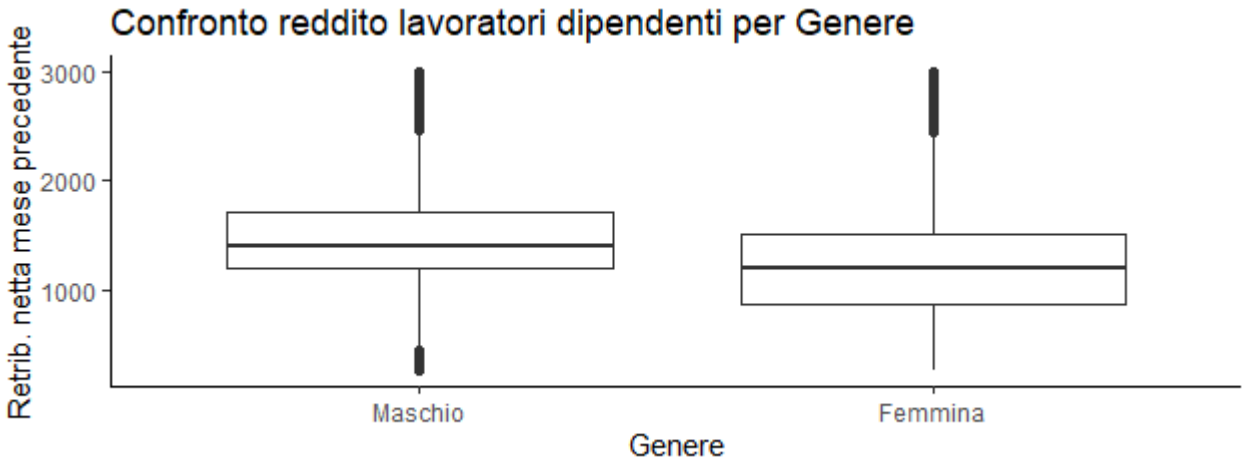

Figura.9: In questa figura si mostrano i box-plot della variabile RETRIC in relazione alla variabile SG11.

Si nota che il reddito mensile risulta generalmente superiore per i lavoratori maschi rispetto alle lavoratrici femmine, assumendo però per i lavoratori maschi una maggiore variabilità rispetto a quello delle lavoratrici femmine.

- *analisi della variabile RETRIC rispetto alla variabile CLETAD (classi d'età)*

CLETAD	Mediana	Media	Dev. Std. Campionaria
15-24 anni	1200	1209.11	449.4
25-34 anni	1330	1340.34	520.6
35-44 anni	1400	1406.13	544.81
45-54 anni	1400	1455.76	566.85
55-64 anni	1330	1366.31	640.309

- Tabella 31

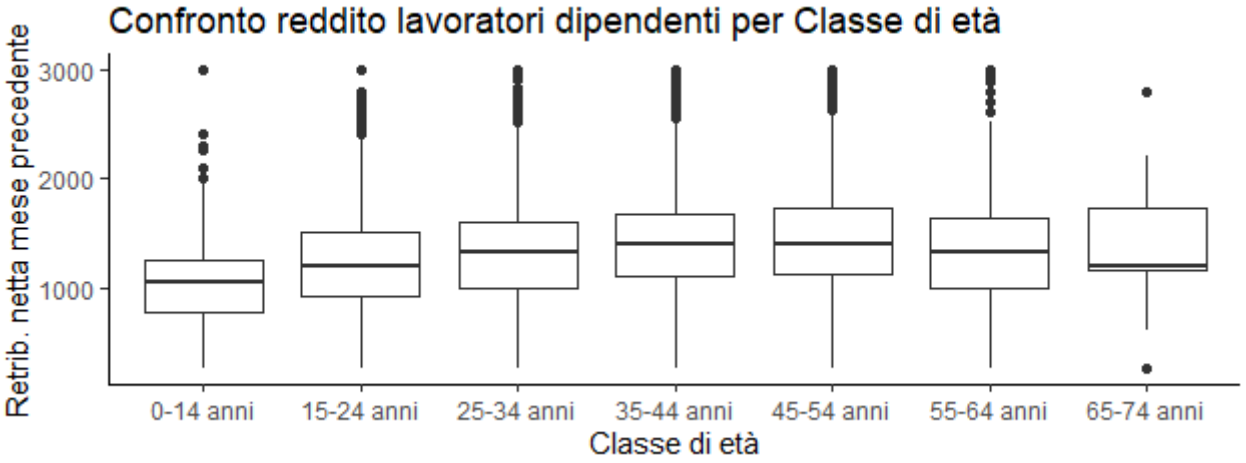

Figura.10: In questa figura si mostrano i box-plot della variabile RETRIC in funzione della variabile CLETAD.

Si nota che, il reddito mensile cresce al crescere delle classi d'età fino alla classe 45-54 anni. riducendosi mediamente per la classe 55-64 anni. Il reddito mensile assume anche una maggiore variabilità al crescere delle classi d'età.

- *analisi della variabile RETRIC rispetto alla variabile EDULEV (titolo di studio)*

EDULEV	Mediana	Media	Dev. Std. Campionaria
Nessun titolo	900	993.15	402.98
Licenza elementare	1055	1024.80	438.65
Licenza media	1200	1175.98	445.37
Diploma 2-3	1290	1257.11	439.67
Diploma 3-5	1330	1338.64	499.63
Laurea	1560	1640.99	607.79

- *Tabella 32*

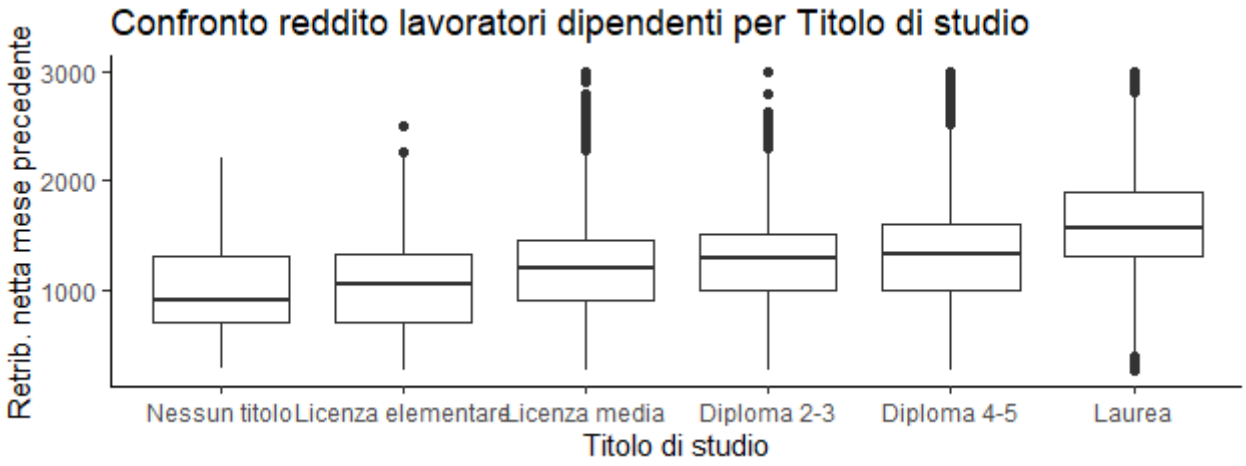

Figura.11: In questa figura si confrontano i box-plot della variabile RETRIC in relazione alla variabile EDULEV.

Si nota che il reddito è fortemente influenzato dal titolo di studio, infatti, con un grado di istruzione superiore aumenta il reddito percepito. Inoltre si evidenzia che, all'aumentare del grado d'istruzione, aumenta anche la variabilità del reddito percepito.

- *analisi della variabile RETRIC rispetto alla variabile C9 (posizione lavorativa*)

C9	Mediana	Media	Dev.Std Campionaria
Dirigente	3000	2648.25	532.47
Quadro	1900	1998.47	546.90
Impiegato	1400	1408.29	452.66
Operaio	1200	1144.95	430.42
Apprendista	990	980.61	315.66
Lavoratore presso proprio domicilio per conto di un'impresa	850	735	270.09

- *Tabella 33*

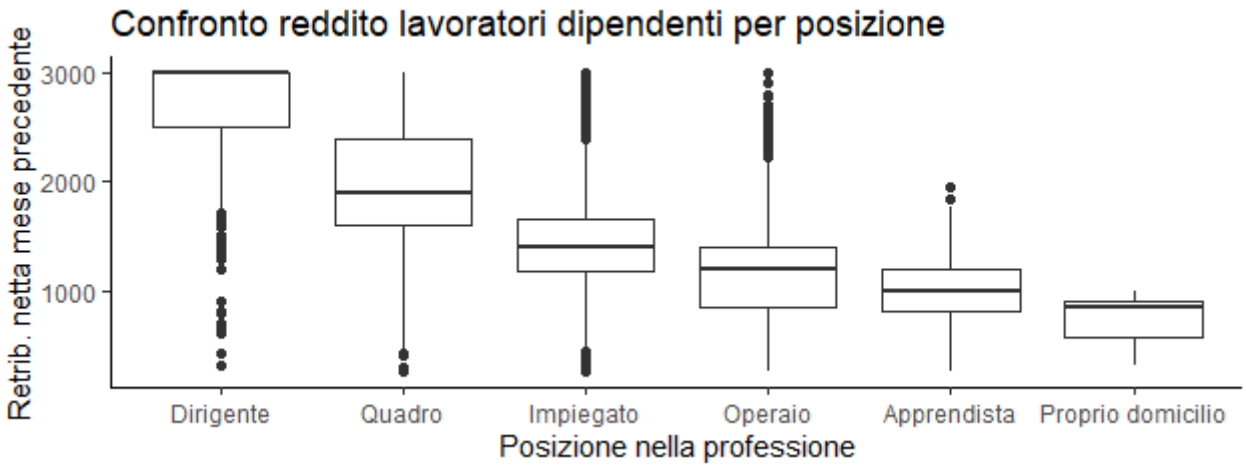

Figura.12: In questa figura si confrontano i box-plot della variabile RETRIC in relazione alla variabile C9.

Si nota che il reddito è fortemente influenzato dalla posizione lavorativa ricoperta, infatti, all'aumentare della posizione lavorativa aumenta il reddito percepito.

- *analisi della variabile RETRIC rispetto alla variabile RIP5 (area geografica)*

RIP5	Mediana	Media
Nord-Ovest	1380	1399.09
Nord-Est	1400	1406.51
Centro	1300	1301.75
Sud	1290	1272.14
Isole	1300	1280.35

- Tabella 34

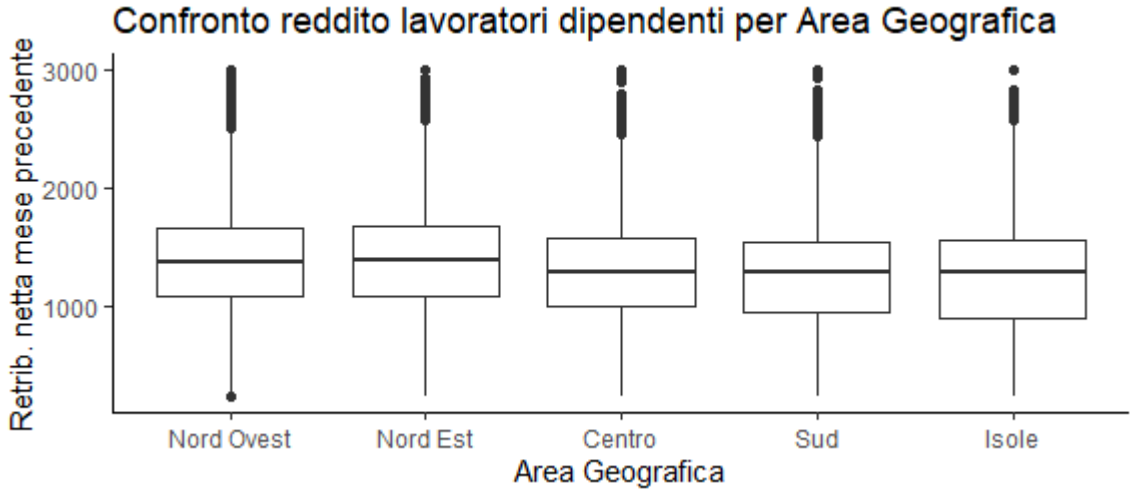

Figura.13: In questa figura si confrontano i box-plot della variabile RETRIC in relazione alla variabile RIP5.

Si nota che lavoratori del *nord Italia* percepiscono un reddito lievemente maggiore rispetto ai lavoratori del *centro,* del *sud* e delle *isole*.

5.3 Test del Chi Quadro

E' stata inoltre effettuata un'analisi di associazione, attraverso il test Chi Quadro, tra alcune coppie di variabili categoriche:

- *titolo di studio e Genere*
- *titolo di studio e Area Geografica*
- *titolo di studio e Posizione professionale*
- *luogo di lavoro e Area geografica*
- *posizione professionale e Genere*

In tutti i casi valutati, è stata verificata la presenza di una dipendenza statisticamente significativa tra alcuni valori assunti dalle due variabili all'interno del dataset di riferimento.

5.4 ANOVA a una via

L'analisi della varianza (ANOVA) è stata effettuata al fine di individuare un'eventuale differenza tra le medie nei gruppi e, quindi, capire se tale differenza fosse relativa ad una effettiva differenza tra le medie nei gruppi.

E' stata applicata l'analisi della varianza alla variabile RETRIC a diverse variabili categoriche: area geografica (RIP5), genere (SG11), livello di istruzione (EDULEV), posizione nella professione (C9).

I valori della statistica *F* ottenuti (rapporto tra la varianza tra i gruppi e la varianza interna ai gruppi) è risultato superiore a 100.

Sono stati riscontrati valori maggiori per le variabili genere (1628, *df = 2*) e Posizione C9 (2161, *df = 5*).

Il valore molto basso per il *p-value* porta sempre al rifiuto dell'ipotesi nulla, secondo la quale tutte le medie sono uguali: ciò non significa che le medie sono tutte significativamente diverse l'una dall'altra, ma che esiste almeno una coppia di medie la cui differenza è statisticamente significativa.

Viene osservato che:

- per quanto riguarda la forma della distribuzione, l'andamento è lo stesso riscontrato a livello generale;
- per quanto riguarda l'omogeneità delle varianze, la distribuzione dei residui generalmente è centrata intorno a valori prossimi allo 0, con distanza interquartile non superiore a 600 e per questo si considera valida l'ipotesi.

5.5. Regressione lineare multipla

Concludendo, viene compiuto uno studio sull'eventuale definizione di un modello di regressione lineare per la variabile che rappresenta la retribuzione RETRIC.

Per fare ciò, vengono prese in considerazione un numero ridotto di variabili rispetto a quelle che possono essere apprezzate nello studio del fenomeno della retribuzione e ciò, con l'obiettivo di individuare le variabili che influenzano maggiormente la variabile in analisi.

Le variabili considerate sono le variabili indipendenti: - genere, - area geografica, - posizione nella professione, - tempo pieno/part-time e - durata del lavoro attuale in mesi

Dalle prove effettuate vengono evidenziati i risultati già riscontrati nelle analisi precedenti: l'analisi dei coefficienti conferma che il genere 'femmina' ha un impatto negativo sulla retribuzione rispetto a 'maschio', così come l'area geografica 'centro o mezzogiorno' rispetto a 'nord'. L'età crescente, invece, influisce positivamente sulla retribuzione.

Per tutti i coefficienti i valori di *p-value* sono molto bassi, a conferma della loro significatività.

Il valore di *R2*, che rappresenta la proporzione di variabilità della variabile RETRIC spiegata dall'insieme di variabili dipendenti considerate, risulta di 0.5437.

Visto che, tale indicatore varia in un intervallo tra 0 e 1 il valore ottenuto è considerato abbastanza soddisfacente considerando la complessità nella modellizzazione del fenomeno in oggetto.

```
Call:
lm(formula = RETRIC ~ SG11 + RIP5 + C9 + C27 + DURATT, data = ds)

Residuals:
     Min       1Q   Median       3Q      Max
-2397.38  -213.01    -8.66   196.81  2288.00

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.099e+03  1.760e+01 176.052   <2e-16 ***
SG11Femmina      -1.879e+02  4.973e+00 -37.791   <2e-16 ***
RIP5Nord Est      2.114e+01  6.159e+00   3.432    6e-04 ***
RIP5Centro       -7.783e+01  6.219e+00 -12.515   <2e-16 ***
RIP5Sud          -1.323e+02  7.256e+00 -18.235   <2e-16 ***
RIP5Isole        -1.460e+02  9.452e+00 -15.446   <2e-16 ***
C9Quadro         -6.215e+02  1.787e+01 -34.778   <2e-16 ***
C9Impiegato      -1.087e+03  1.602e+01 -67.844   <2e-16 ***
C9Operaio        -1.345e+03  1.607e+01 -83.695   <2e-16 ***
C9Apprendista    -1.478e+03  3.158e+01 -46.798   <2e-16 ***
C9Proprio domicilio -1.697e+03  1.500e+02 -11.315   <2e-16 ***
C27              -4.728e+02  6.149e+00 -76.898   <2e-16 ***
DURATT            6.369e-01  1.747e-02  36.469   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 365.4 on 25779 degrees of freedom
Multiple R-squared:  0.5376,    Adjusted R-squared:  0.5374
F-statistic:  2498 on 12 and 25779 DF,  p-value: < 2.2e-16
```

Figura.14: In questa figura si espongono i risultati ottenuti per il modello di regressione multipla della variabile RETRIC utilizzando le variabili SG11, RIP5, C9, C27, DURATT.

Viene calcolato il VIF (fattore di inflazione della varianza) per ciascuna variabile esplicativa:

Variabile	VIF
SG11	1.19
RIP5	1
C9	1.14
C27	1.16
DURATT	1.08

- *Tabella 35*

Osservando i valori è possibile concludere che non vi è presenza di multicollinearità tra le variabili esplicative.

Conclusioni

Le analisi svolte in questo lavoro hanno permesso di valutare la situazione lavorativa in Italia nel secondo trimestre del 2020.

Dai risultati ottenuti è possibile evidenziare un aumento della retribuzione al crescere dell'età e del titolo di studio ottenuto, ma anche purtroppo una differenza retributiva in base al genere.

Dall'analisi bivariata della retribuzione in relazione al genere, emerge infatti, che le donne guadagnino in media meno degli uomini.

La retribuzione invece non varia molto spostandosi da un'area geografica ad un altra come evidenziato dai box-plot presentati nella *Figura.13*

La variabile RETRIC risulta quindi influenzata fortemente dalla posizione lavorativa e dalla classe d'età.

L'analisi del tasso di occupazione e del tasso di attività evidenzia anche in questo caso una disparità dovuta al genere, in particolare una situazione negativa per il gruppo "Femmine".

Dallo studio effettuato sulla regressione multipla si evidenziano infatti i risultati già riscontrati nelle analisi precedenti: corroborando attraverso l'analisi dei coefficienti l'ipotesi che il genere 'femmina' abbia un impatto negativo sulla retribuzione rispetto a 'maschio', l'età crescente, così come la posizione lavorativa invece, influiscono positivamente sulla retribuzione.

Infine l'analisi del fattore d'impatto d'inflazione della varianza (VIF) ci ha permesso di verificare l'assenza di multicollinearità nel nostro modello di regressione multipla.

Sitografia

Rilevazione sulle forze di lavoro, Microdati Istat - https://www.istat.it/it/archivio/127792