Università degli studi di Milano-Bicocca

Big data in Business, economics and society

Progetto Finale

FEDERLEGNO

Analisi della sensibilità degli associati Federlegno a topic attuali

---

Autori:

Beatrice Fumagalli - 784549 - b.fumagalli9@campus.unimib.it

Nicholas Missineo – 791050 – n.missineo@campus.unimib.it

Francesco Simoncelli - 834313 – f.simoncelli@campus.unimib.it

Beatrice Somaschini – 789554 - b.somaschini1@campus.unimib.it

Introduzione

Il termine **Industria 4.0** è ormai noto a quasi tutte le aziende italiane, anche nel settore arredo. In termini generici, Industria 4.0 sta ad indicare una generale tendenza alla digitalizzazione: proprio per questo si parla anche di **quarta rivoluzione industriale**.
Si tratta non solo dell'automatizzazione dei processi industriali, ma anche della gestione di una grande mole di dati, i cosiddetti *Big Data*, della fruizione di questi dati per formulare analisi corrette e di valore per l'azienda, dell'interconnessione fra macchina e uomo.
Questa nuova tendenza si è concretizzata in Italia in una serie di misure e di agevolazioni volte a incentivare la digitalizzazione delle aziende italiane, il piano nazionale industria 4.0.

Il cambiamento sostanziale nel **settore legno arredo** è stato il passaggio da una produzione seriale e standardizzata, alla cosiddetta ***mass customization***, ovvero la "personalizzazione di massa"; infatti, i clienti con potere di acquisto hanno richiesto via via un grado di personalizzazione del prodotto sempre maggiore: una modifica rispetto al prodotto da catalogo o anche la progettazione di un nuovo prodotto, *ad hoc.*
Innovazione significa crescita e alle aziende italiane del settore dell'arredo questo non è sfuggito. Gli incentivi di industria 4.0 sono stati colti e hanno portato alla digitalizzazione e all'innovazione tecnologica; gli effetti positivi sono evidenti all'interno del mercato e i dati parlano chiaramente[1], ma ciò che non deve essere dimenticato è che questi effetti positivi devono essere comunicati al cliente e non solo, in quanto sono le persone che fanno l'azienda e proprio per questo si parla di **rivoluzione culturale**, ancora prima che di strumenti.

L'azienda del futuro nell'immaginario collettivo non solo è un'azienda innovativa, automatizzata e *smart*, ma è anche un'azienda in cui si prendono scelte più sostenibili rispetto al passato, nei riguardi dell'ambiente e anche in direzione di una sostenibilità sociale. Un'industria sempre connessa, vicina al cliente ma anche ai suoi dipendenti e fornitori.
Con questo progetto si vuole analizzare il livello di sensibilità rispetto a queste tematiche delle aziende del settore legno arredo e come questa sensibilità venga trasmessa al pubblico, se in maniera efficiente o non.
Le principali domande a cui si è cercato di trovare una risposta sono state:

- Come si sta muovendo l'industria del legno in merito ai temi di innovazione, digitalizzazione e *smartness*?
- Quanto le scelte effettuate da tali aziende sono ecosostenibili?
- Il tema 'design dei prodotti' rimane centrale nella *mission* delle aziende del legno?

Per condurre questa analisi sono state prese in considerazione le aziende associate alla famosa federazione **FederlegnoArredo[2],** cuore della filiera italiana del legno-arredo. Di un campione di queste aziende sono state poi estrapolate le due pagine "vetrina" del loro sito *web*, ossia l'*homepage* e la sezione "chi siamo" per analizzare, attraverso tecniche di ***Text Mining***, la sensibilità dell'azienda in questione rispetto alle principali tematiche dell'Industria 4.0: **innovazione, sostenibilità, design** e **social network**. Delle aziende prese in esame è

[1] https://www.ilsole24ore.com/art/impresa-e-territori/2018-03-16/l-arredo-investe-ricerca-e-produzione-40-e-ritrova-crescita-114618.shtml?uuid=AEe844HE
[2] http://www.federlegnoarredo.it/

stato poi fornito un ***ranking*** in grado di mostrare in quale misura quest'ultime sono sensibili ai temi salienti presentati in precedenza.

Workflow

Di seguito viene presentato sinteticamente il flusso di lavoro che è stato seguito al fine di analizzare la sensibilità degli associati Federlegno ai diversi *topic*, presentati di seguito.

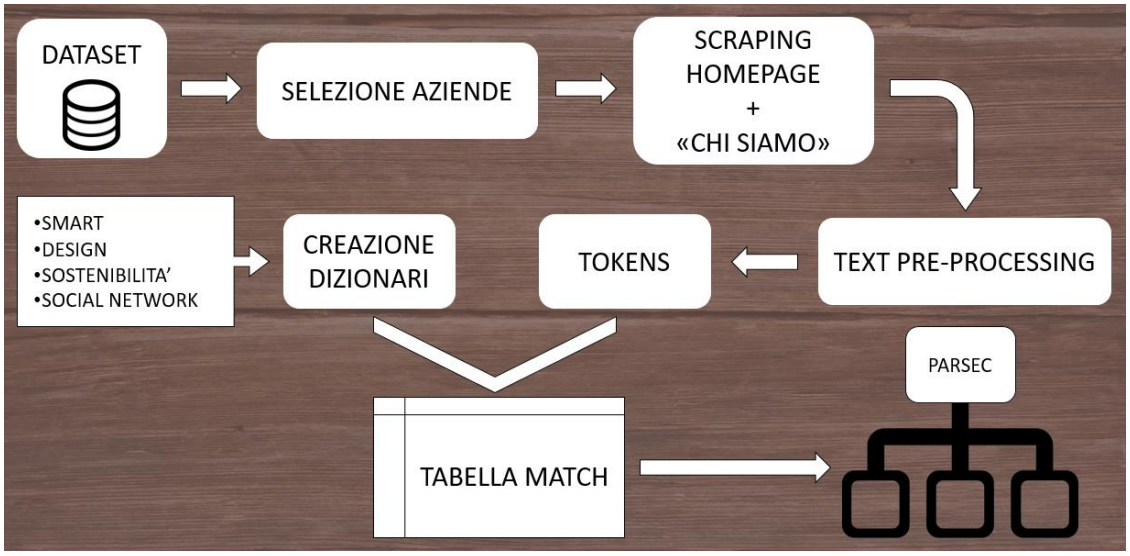

Figura 1 - Workflow del progetto

1) Esplorazione iniziale del dataset
2) Selezione di un campione di aziende
3) *Scraping* della *homepage* e della sezione "chi siamo" di ogni azienda
4) *Text pre-processing* per ottenere dei *tokens* puliti
5) Creazione di quattro dizionari, uno per ciascun *topic*
6) *Matching* tra i dizionari creati e il testo pulito
7) Utilizzo di *Parsec* per il *ranking*

Il Dataset

Il dataset fornitoci ('dataset.csv') è composto da 2.022 aziende e da 6 attributi:

- **Ragione sociale:** nome dell'azienda;
- **Partita IVA:** partita IVA aziendale;
- **Provincia:** provincia della sede aziendale;
- **Indirizzo sede legale Regione:** regione della sede aziendale;
- **Ultimo Bilancio Ricav delle vendite:** fatturato aziendale;
- **SISTEMA PREVALENTE:** indica il sistema all'interno del quale l'azienda produce;

Selezione del campione di aziende

Prima di poter selezionare il campione di aziende sul quale eseguire le analisi, è stato necessario suddividere quest'ultime in quattro dimensioni in base al fatturato[3], ottenendo così la seguente classificazione:

- MICRO: fatturato inferiore a 2 milioni, 774 aziende
- PICCOLE: fatturato compreso tra 2.000.000 e 10.000.000, 834 aziende
- MEDIE: fatturato compreso tra 10.000.000 e 50.000.000 , 340 aziende
- GRANDI: fatturato oltre i 50 milioni, 74 aziende

Successivamente sono state selezionate le aziende della Regione Lombardia, in particolare quelle "brianzole" localizzate in provincia di Monza e Brianza e sono state prese come campione le aziende di dimensione media e grande, appartenenti ai seguenti sistemi: arredamento, illuminazione e ufficio.

Scraping

Si è scelto di eseguire lo *scraping* solamente delle due pagine "vetrina" del sito web di ciascuna azienda, ossia l'*homepage* e la sezione "chi siamo". Lo *scraping* è stato eseguito attraverso l'utilizzo della funzione ***'read_html'*** e delle due librerie ***tidyverse*** e ***rvest*** messe a disposizione da ***R***. È stato ritenuto opportuno eseguire lo *scraping* sulle pagine *web* in lingua inglese, per permettere un accurato *text pre-processing* per le successive analisi.

Text pre-processing

La fase di *pre-processing* del testo è stata articolata in un'unica funzione chiamata *'text_preproc'* e successivamente applicata al testo di entrambe le sezioni, *homepage* e "chi siamo", utilizzando la libreria ***Gensim*** messa a disposizione da ***Python***.
La funzione svolge i seguenti compiti:

- Rimozione punteggiatura
- Rimozione righe vuote
- Rimozione dei numeri
- Stop *words*: rimozione delle parole brevi prive di significato semantico (es. articoli, preposizioni, ecc.)
- Rimozione *tags*
- *Stemming*: processo di riduzione della forma flessa di una parola alla sua forma radice
- *Tokenization:* divisione del testo in singole parole, *tokens.*

Established in 1954, Zanotta is one of the recognized leaders in Italian industrial design since ever.						
establish	zanotta	recogn	leader	italian	industri	design

Figura 2 - Esempio text pre-processing

[3] http://www.economiaziendale.net/lezioni/azienda/aziende_piccole_medie_grandi.htm

Creazione dizionari

Sono stati creati quattro dizionari, uno per ciascuno dei seguenti *topic*:

- ***Smart*** – Innovazione
- ***Design***
- ***Sustainability*** – Sostenibilità
- ***Social Network***

La creazione dei dizionari è stata graduale. In principio sono state scelte le parole che sembravano essere più affini ai *macro-topic* scelti, successivamente quest'ultime sono state integrate con delle nuove, derivanti da un'analisi più approfondita dei testi delle diverse aziende.

Figura 3 - Dizionario topic Smart

Figura 4 - Dizionario topic Design

Figura 5 - Dizionario topic Sustainability

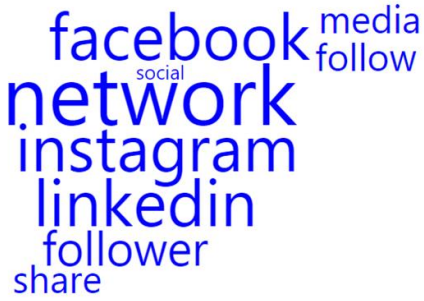

Figura 6 - Dizionario topic Social Network

Alle parole contenute in ciascun dizionario è stato eseguito il processo di *Stemming,* al fine di consentire il successivo *matching* con i *tokens* puliti derivanti dalla precedente fase di *pre-processing* del testo estrapolato dai siti *web*.

Matching

Il *matching* tra i dizionari sopra descritti e i *tokens* puliti di ciascuna azienda è stato eseguito attraverso una funzione in *Python* appositamente creata, al fine non solo di stabilire all'interno del testo di quale azienda quali parole di quale *topic* comparissero, ma anche per contare il numero di occorrenze di ciascuna parola del dizionario all'interno del testo di ciascuna azienda. Si è così ottenuto un *DataFrame* finale *'tab_finale_parsec.csv'* contenente:

- **Azienda:** nome dell'azienda
- **Sistema:** arredamento, illuminazione o ufficio
- **Dimensione:** media o grande
- **Smart_count:** numero di occorrenze delle parole del dizionario del *topic Smart* all'interno del testo
- **Design_count:** numero di occorrenze delle parole del dizionario del *topic Design* all'interno del testo
- **Sust_count:** numero di occorrenze delle parole del dizionario del *topic Sustainability* all'interno del testo
- **Soc_net_count:** numero di occorrenze delle parole del dizionario del *topic Social Network* all'interno del testo

Ranking

Partendo dalle informazioni contenute nel DataFrame sopra descritto si è quindi deciso di procedere con un *ranking* delle aziende basato sulla sensibilità di quest'ultime rispetto ai temi trattati da ciascun *topic*, sensibilità misurata in base alla frequenza delle occorrenze.
Il *ranking* è stato calcolato utilizzando ***R***, attraverso la libreria ***Parsec***. Del *DataFrame* iniziale sono state considerate solo le variabili numeriche ordinali del conteggio delle occorrenze, ossia *Smart_count*, *Design_count*, *Sust_count* e *Soc_net_count* e di queste ne sono state estratte i profili. A questo punto è stata creata la matrice **MPR**: ***mutual ranking probability***. Quest'ultima è stata scomposta attraverso decomposizione spettrale e la prima colonna della matrice degli autovettori è stata utilizzata per dare un ordinamento ai profili, da cui si è poi potuto risalire all'ordinamento delle diverse aziende.
Si precisa che si è deciso di considerare ciascuno dei quattro *topic* con lo stesso peso, tuttavia se dovesse ritenersi opportuno è possibile dare maggior peso ad alcune variabili piuttosto che ad altre.

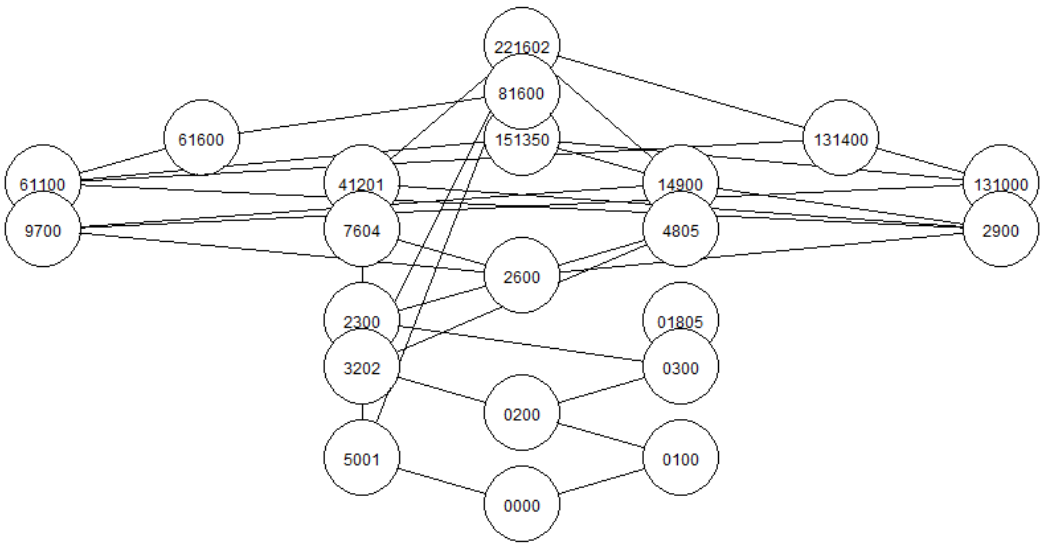

Figura 7 – diagramma di Hasse

La figura 7 rappresenta il diagramma di Hasse ottenuto attraverso l'operazione di *ranking*. Il diagramma mostra un ordinamento delle combinazioni presenti all'interno del *dataset* analizzato. Ogni nodo rappresenta una delle combinazioni presenti nel *dataset*, da notare come sono state considerate unicamente le combinazioni presenti e non quelle possibili. Sebbene risulti difficile da visualizzare è possibile avere un'idea dell'ordinamento esistente. Questa visualizzazione risulta sempre più impraticabile al crescere del numero dei profili.

Risultati

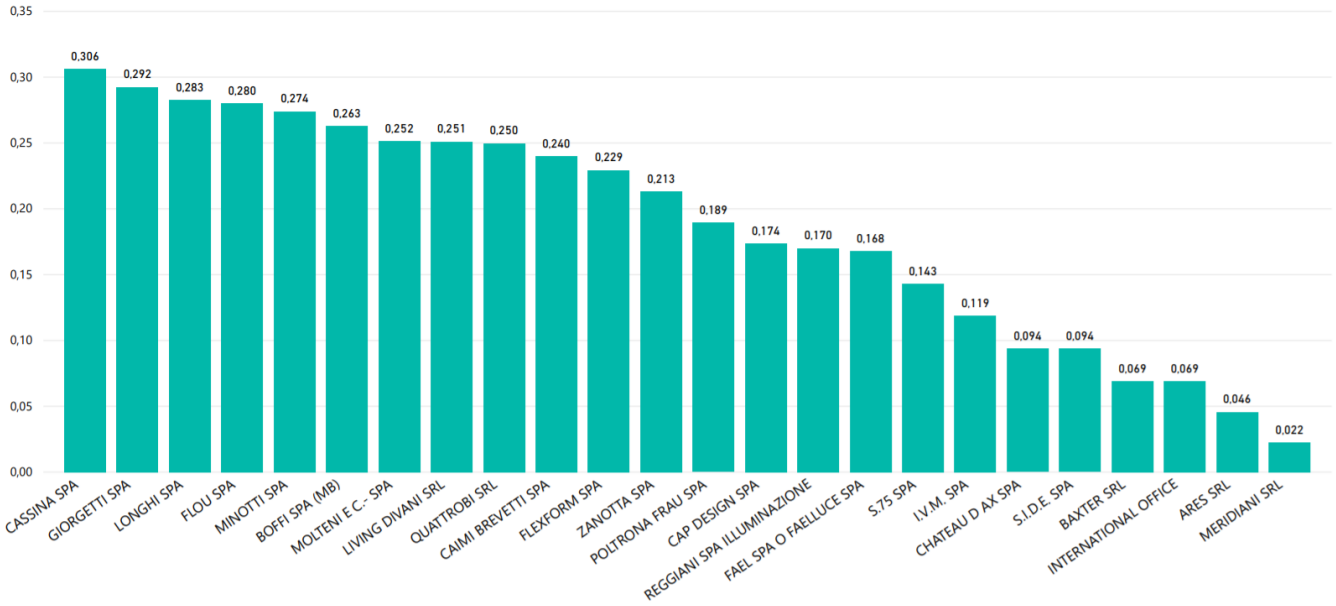

Figura 8 - Bar chart Ranking

Come si può notare dal ***Bar chart*** sopra raffigurato, l'azienda che risulta essere il più sensibile alle tematiche racchiuse nei quattro *topic* (Innovazione, Design, Sostenibilità e Social Network) con un *rank* dello 0,306 è **Cassina S.P.A.**, una grande azienda brianzola del sistema arredamento. Nonostante il fatturato di gran lunga inferiore alla prima sopra citata, al secondo posto si trova **Giorgetti S.P.A.**, con uno *score* di 0,292, la quale è anche l'unica azienda che è risultata essere sensibile alla tematica della sostenibilità.
Il terzo e quarto posto vedono ancora due aziende di medie dimensioni: **Longhi S.P.A.** e **Flou S.P.A.,** mentre al quinto posto si classifica un'azienda di grandi dimensioni**: Minotti S.P.A.**.
L'azienda meno sensibile alle tematiche trattate dai *topic* risulta essere **Meridiani S.R.L.**, azienda media brianzola che, tra le aziende prese in analisi, risulta essere anche l'unica non sensibile a nessuno dei quattro *topic* trattati.

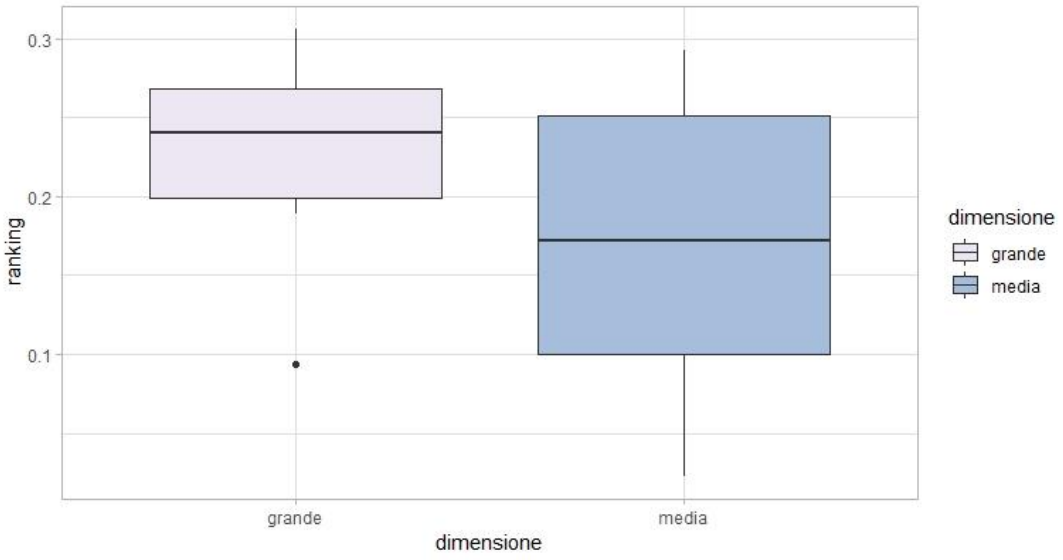

Figura 9 - Box plot Ranking per dimensione

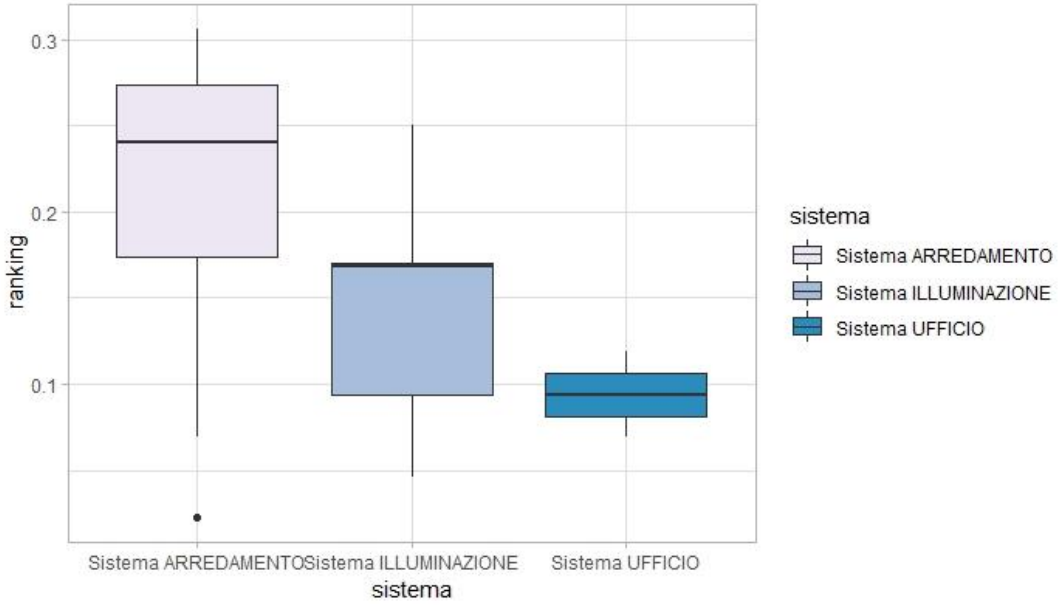

Figura 10 - Box plot Ranking per sistema

In linea generale, come si può notare dal primo *Box plot*, le aziende di grande dimensione, ossia con un fatturato più elevato, tendono ad avere un *rank score* più elevato e pertanto risultano essere più sensibili alle tematiche di Innovazione, Design, Sostenibilità e Social Network rispetto alle aziende medie. Così come le aziende del sistema arredamento risultano avere una media di *rank score* più elevato rispetto agli altri due sistemi riconosciuti dalla federazione FederlegnoArredo, ossia illuminazione e ufficio.

Risultati – Test Kruskal-Wallis

Per poter approfondire a livello statistico le intuizioni avute da una prima ispezione grafica si può procedere attraverso un test statistico. In questo caso, dato che si è interessati a testare se le differenze di una variabile continua suddivisa tra gruppi risultano essere significative, un test statistico adeguato è il test Kruskal-Wallis. In Statistica, il test di Kruskal-Wallis è un metodo non parametrico per verificare l'uguaglianza delle mediane di diversi gruppi, cioè per verificare che tali gruppi provengano da una stessa popolazione (o da popolazioni con uguale mediana); è il corrispondente non parametrico dell'analisi di varianza (ANOVA) in cui i dati vengono sostituiti dal loro rango e viene solitamente usato quando non può essere assunta una distribuzione normale della variabile oggetto di analisi.
Il test è stato calcolato su **R** attraverso la funzione '*kruskal.test'*. Si riportano di seguito i risultati ottenuti:

```
        Kruskal-Wallis rank sum test

data:  xs
Kruskal-Wallis chi-squared = 5.1072, df = 2, p-value = 0.0778
```

Figura 11- test Kruskal-Wallis per la significatività del SISTEMA

```
        Kruskal-Wallis rank sum test

data:  x
Kruskal-Wallis chi-squared = 1.5224, df = 1, p-value = 0.2173
```

Figura 12- test Kruskal-Wallis per la significatività della DIMENSIONE

Il primo test effettuato *(Figura 11*) va a verificare se il valore del *ranking* si modifica tra i diversi gruppi di aziende clusterizzate in base al sistema di produzione prevalente. Il valore del p-value risulta essere 0.0778, quindi ad un livello di significatività del 5%, si va ad accettare l'ipotesi nulla, cioè si va ad accettare che non vi siano differenze significative tra le mediane di *ranking* all'interno dei diversi gruppi. Si potrebbe andare a rifiutare il test e quindi ammettere differenze di *ranking*, ad un livello di significatività del 10%. Pertanto nonostante dal Box Plot emergeva una differenza tra aziende appartenenti ai diversi sistemi, questa differenza viene smentita dal test statistico.
Nel secondo caso *(Figura 12)* il test effettuato sulle aziende clusterizzate in base alla dimensione viene invece accettato a qualsiasi livello di significatività, poiché il p-value di 0,2173 supera la soglia di significatività del 10%. Quindi sebbene dal *Box plot* poteva apparire una differenza tra aziende medie ed aziende grandi, il test statistico nega questa differenza non trovando una discriminazione.
Dal momento che il test scelto non richiede una numerosità uguale delle distribuzioni, ma è preferibile che essa sia simile ed inoltre richiede un campione consistente, va tenuto conto sia dell'esiguità del campione totale di aziende considerato ed, in particolare, della differente numerosità campionaria nei vari gruppi analizzati.

In conclusione si può dire che le diverse aziende prese in esame riportano differenze sostanziali circa la sensibilità ai *topic* trattati, se prese singolarmente. Questa differenza di comportamento non è però spiegabile in termini di dimensione e sistema di appartenenza.

Sviluppi futuri

Il principale sviluppo futuro sarebbe quello di estendere l'analisi descritta in questo *report* a tutte le aziende associate alla federazione FederlegnoArredo. Ciò non è stato possibile principalmente per una questione di tempistiche e pertanto si è preferito privilegiare lo sviluppo di un metodo che possa essere ripetuto estendendo il campione. Il secondo impedimento è stata la mancanza di uno strumento di *Scraping* che permettesse di estrarre il contenuto di tutte le pagine del sito *web* di ciascuna azienda, vincolando così l'analisi alle due pagine 'vetrina' del sito, ossia *homepage* e sezione "chi siamo".