AmesHousing Analysis

Davide Mascolo

12 Gennaio 2021

Presentazione del problema

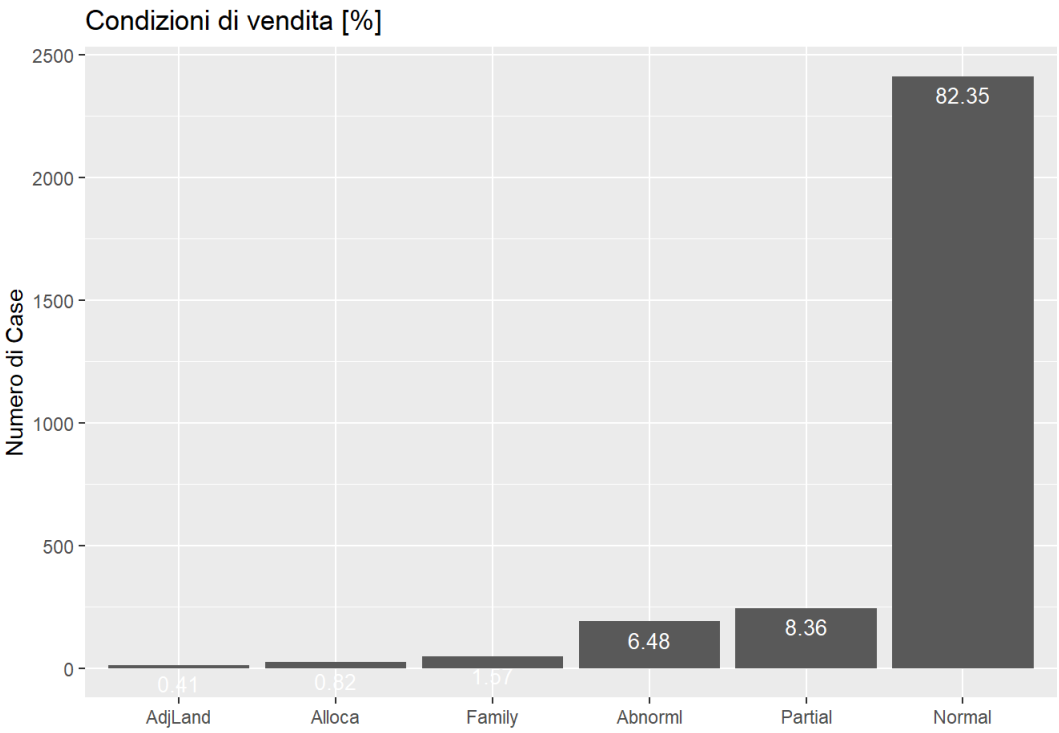
Si vuole prevedere il prezzo di alcune case in vendita ad Ames, Iowa.

Informazioni sui dati

Il set di dati di Ames Housing contiene 80 variabili

Per maggiori informazioni clicca qui.

Step 1 - Data Wrangling



La prima fase riguarda la pulizia dei dati, partendo da raw data. Le operazioni effettuate sono state le seguenti:

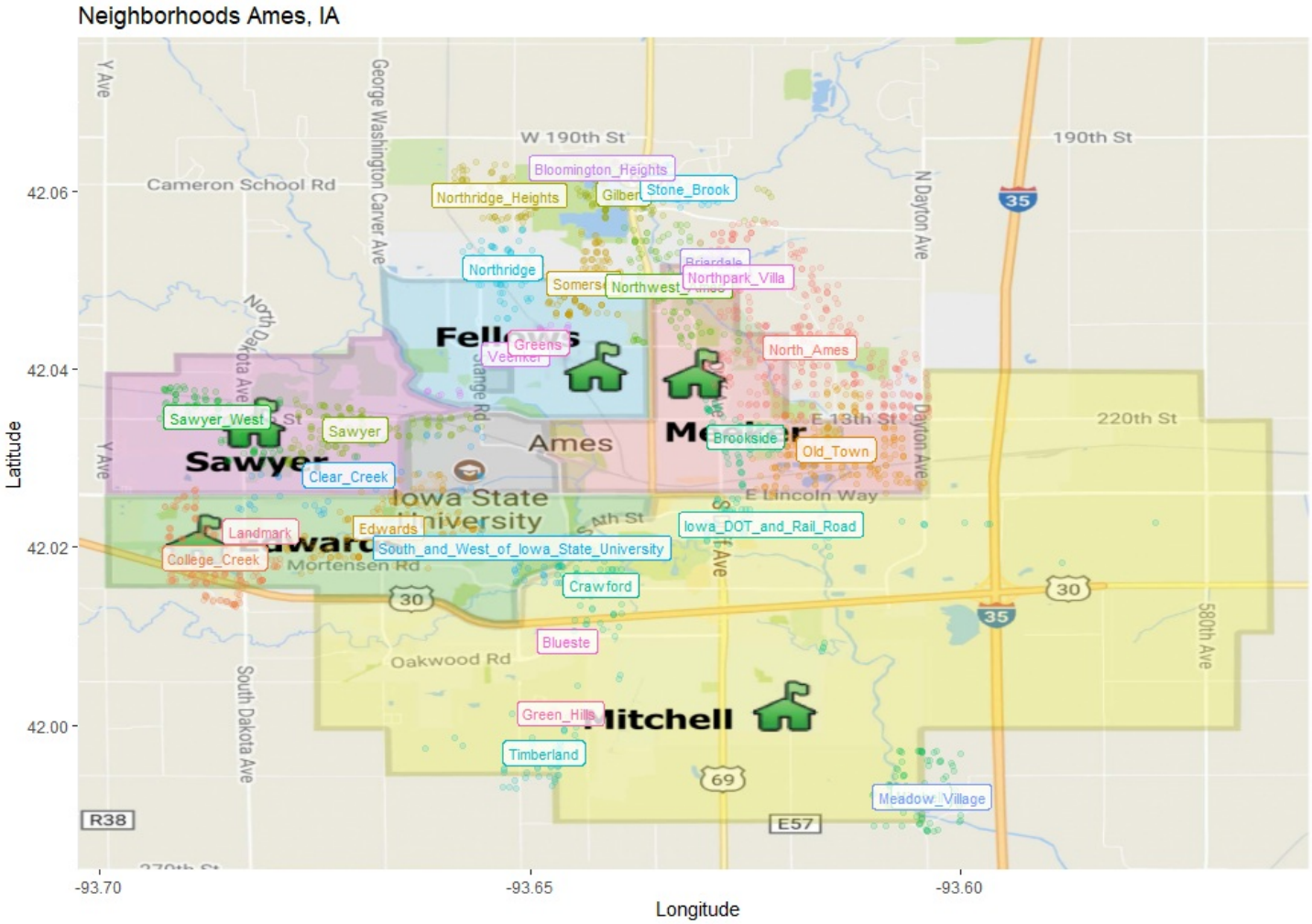
- Tutti i **fattori** sono stati **ordinati.**
- **PID ed Order** vengono **rimossi.**
- **Gli spazi ed i caratteri speciali** nei nomi delle variabili **vengono modificati.** Ad esempio, **SalePrice diventa Sale_Price.**
- Dove possibile, **molti valori mancanti sono stati ripristinati**, ad esempio con la variabile **No_Basement.**
- In altri casi, **variabili conteneti troppi valori mancanti sono state rimosse(Garage_Yr_Blt).**
- Concentrandoci sulla variabile **Sale_Condition**, che indica la **condizione di vendita** di un alloggio, prendiamo in considerazione solo le **vendite "Normal", che sono l'82%.** delle vendite registrate nel campione.

- Le **altre tipologie** di vendite riguardano **vendite parziali, pignoramenti, case ereditate e vendite allo scoperto,** sulle quali andrebbe fatto un **discorso diverso** per prevedene il prezzo in quanto **seguono modelli diversi dalle vendite normali.**

Step 2 - EDA

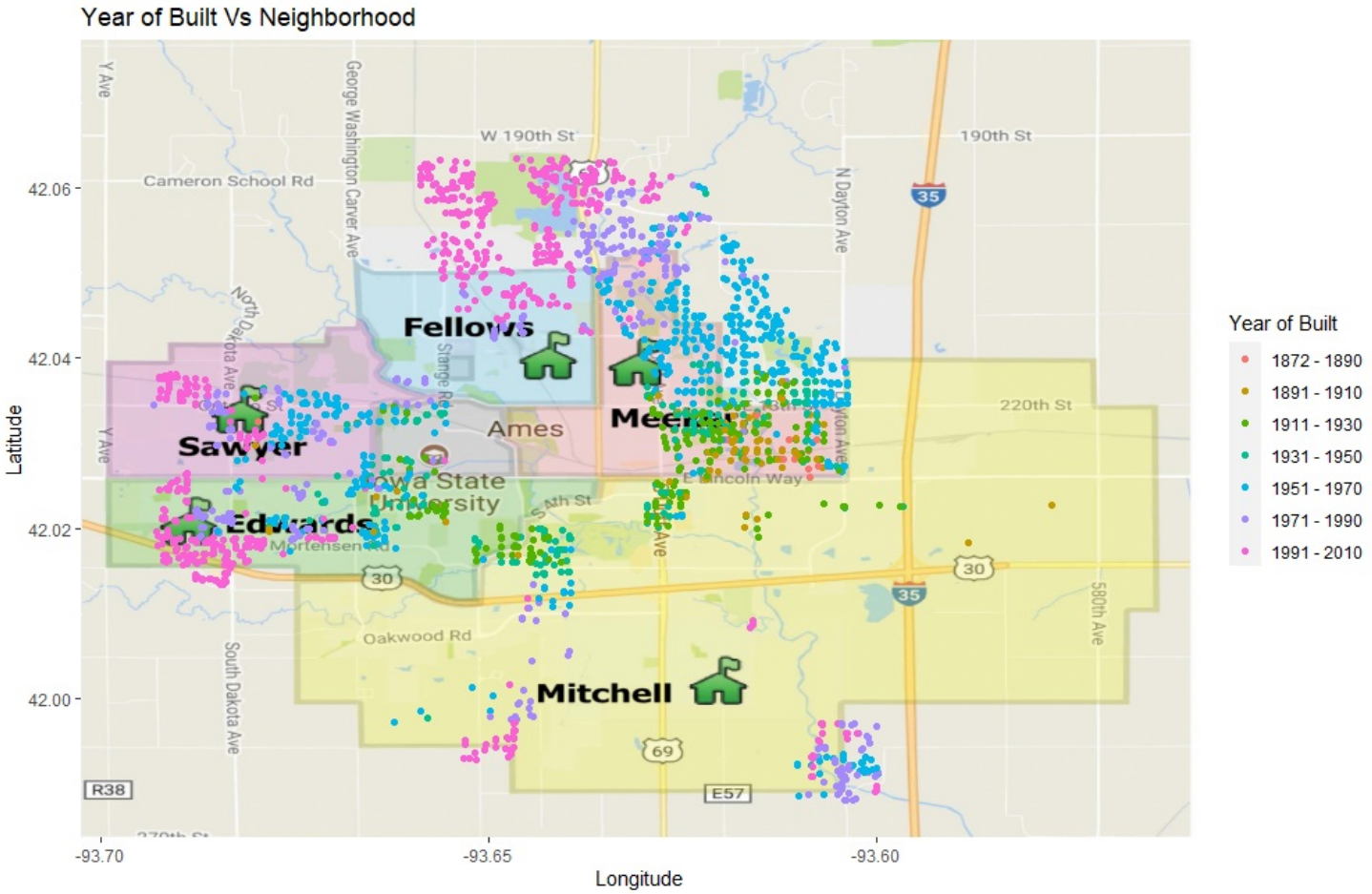
Analisi Spaziale e Temporale

Spazio



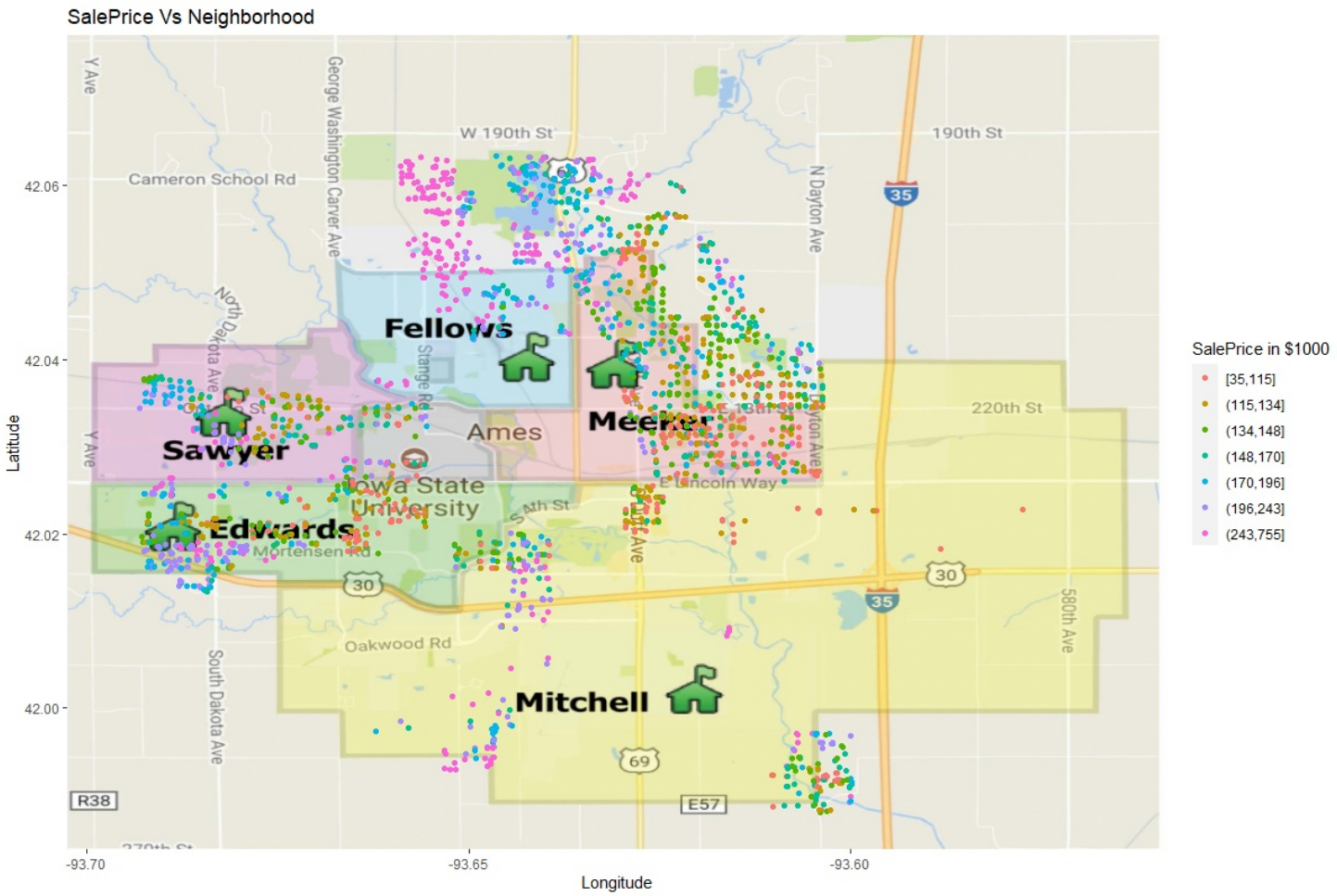
- Ames si trova lungo il **confine occidentale della Story County**, vicino all'incrocio tra **l'Interstate 35 e la U.S. Route 30.** Un'altra strada, ovvero la **U.S. Route 69**, attraversa Ames. Non solo le strade attraversano Ames, ma anche due piccoli corsi d'acqua, il **South Skunk River ed il Dquaw Creek.**
- Questa prima analisi, aiuta a focalizzarsi meglio sul problema, dividendo Ames in cinque distretti: **Fellows, Meeker, Mitchell, Edwards e Sawyer.**
- **Fellows** risulta il distretto maggiormente esposto a **Nord**, dove ci sono quartieri come **Northridge** e la rispettiva zona alta, **Greens e Veenker** nella zona centrale, **Somerset**, ed al confine troviamo la zona alta di **Bloomington** ed il quartiere **Gilbert**.
- Verso **est** troviamo il distretto del **Meeker**, che si divide tra **Old Town e Brookside nella zona bassa**, **Nord Ames** a Nord, ed al **confine con Fellows** troviamo **Stone Brook, Northpark Villa e Briardale**, con il quartiere Nord-Ovest di Ames che segna in maniera importante il confine tra i distretti Fellows e Meeker.
- **Mitchell** risulta il distretto con **maggiore estensione territoriale**; nonostante questo dato, abbiamo, nel nostro campione, **pochi alloggi** che appartengono a questa regione. Gli unici, infatti, si trovano nel **Meadow Village**, nella zona maggiormente esposta a **sud-est**. **Altri** alloggi li troviamo verso il quartiere **Timberland**, nel **centro-sud** del Mitchell ed un altro gruppo di case si vedono al confine tra il quartiere **Crawford e la zona a sud della Iowa State University.**
- Proprio la **State University** rappresenta il **punto di contatto** tra il distretto del **Sawyer ed Edwards**, in quanto le case che circondano questo edificio storico, si trovano principalmente in questi due distretti

Quali zone sono state costruite recentemente?



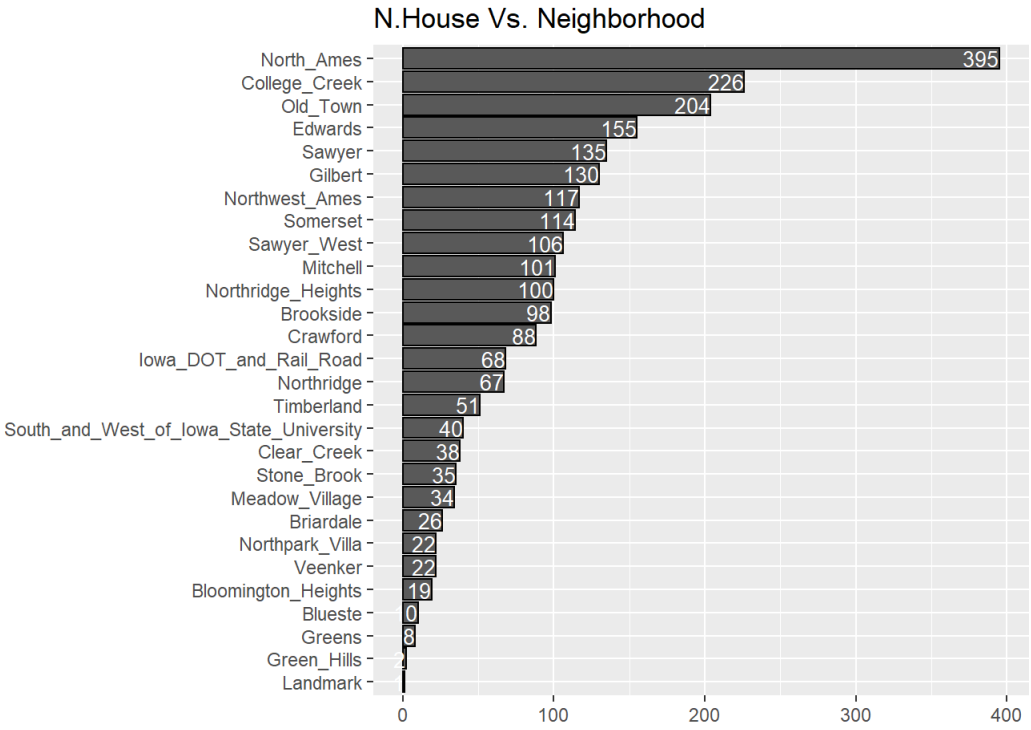
- Le case costruite tra il **1872 ed il 1890**, si trovano nella zona **sud-est del distretto Meeker**.
- Anche le case costruite tra il **1891 ed il 1910**, si trovano nello **stesso distretto** delle case costruite nel decennio prima. Notiamo che nella **zona universitaria**, abbiamo qualche immobile che risale a fine **'800/inizio '900** e questo, data la presenza della State University nei dintorni, ci fa pensare che si tratti di **stabilimenti universitari**.
- Spostandoci verso il **centro-est** troviamo quelle case il cui anno di costruzione risale agli anni tra il **1911 ed il 1930**.
- Continuando verso i **confini di Ames**, troviamo le case che rispettivamente sono state costruite tra il **1931 ed il 1950**, maggiormente nella zona a **nord-est del distretto Meeker**, ed a **sud-ovest della State University**. Sempre in questa zona, fissando come riferimento lo stabilimento universitario, troviamo le case costruite tra il **1971 ed il 1990** ed ancora **verso il confine** abbiamo le case con **costruzione meno datate** ovvero risalenti al **(1991;2010]**.
- La **struttura di questi dati temporali**, indica che effettivamente sul nostro campione, **la costruzione delle case ha avuto come punto di partenza il sud-Meeker per poi spostarsi verso l'esterno.** Infine, **a partire dal 1950, sono iniziate le costruzioni nella zona sud-est del Mitchell.**

Prezzi delle case per ogni quartiere.



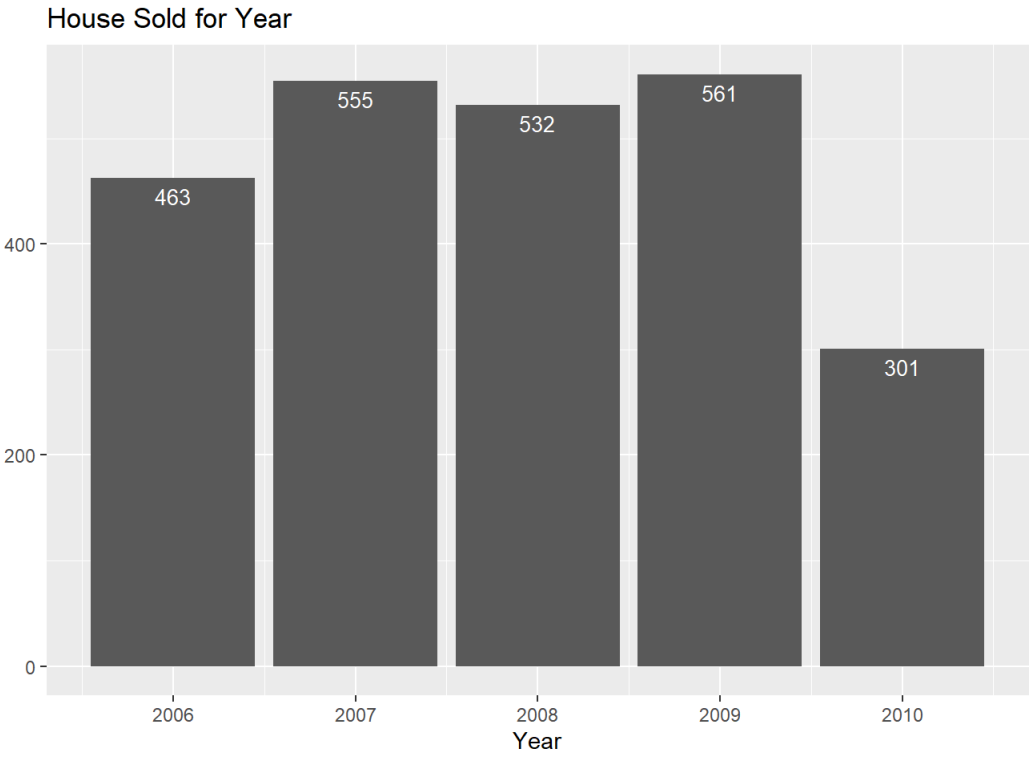
- **Quartieri diversi, hanno prezzo diversi**, sia in termini di **cifre** che di **tenuta del prezzo stesso**.
- Il distretto del **Nord-Fellows**, ad esempio, oltre ad essere un distretto **giovane** per l'anno di costruzione, si dimostra anche un distretto con **prezzi maggiori**. Questo **dipende proprio dall'anno di costruzione** delle abitazioni, ma probabilmente anche dalla vicinanza al famoso club, **Ames Country Club.**
- **Nel distretto di Sawyer**, ci sono **molte case costruite tra il 1995 ed il 2010**, quindi recenti rispetto alle altre case contenute nel campione. Questo farebbe pensare che il prezzo delle case sia alto in questo distretto, ma qui possiamo fare una divisione in due sub-regioni. Si vede che verso il confine ad Ovest del Sawyer, abbiamo un gruppetto importante di case con prezzi tra i **170.000$ ed i 196.000$.**
- Spostandoci verso la State University, invece, aumentano le case con prezzi tra i **115.000$ ed i 134.000$** ed altre ancora con prezzi tra i **35.000$ ed i 114.000$.** Questo aspetto trova coerenza con il fenomeno sociale di Ames, ovvero che la zona universitaria ha una consistenza abitativa importante, che ospita molti appartamenti per studenti, locali notturni, ristoranti ed altri stabilimenti unici di Ames.
- Infine, risulta interessante notare cosa accade a **Sud del distretto del Mitchell**, dove nonostante ci siano **molte case costruite recentemente, i prezzi** nella maggior parte delle abitazioni sono nella **fascia medio-bassa,** probabilmente per la troppa vicinanza all' **Ames Municipal Airport**, collocato proprio a ridosso dell'intersezione tra la **U.S. Route 30 e la U.S. Route 69.**

Numero di case vendute per ogni quartiere

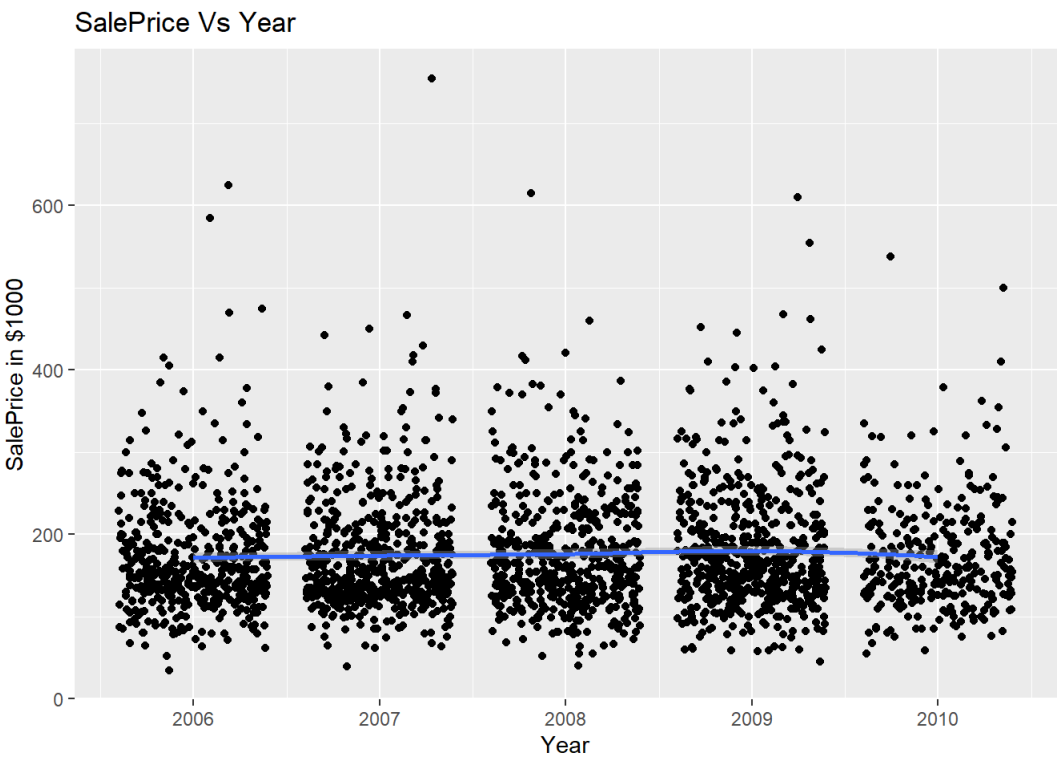


- Il distretto in cui si registra il numero di case **maggiormente vendute** risulta **North Ames**, ovvero la zona a **Nord del Meeker.**
- Come abbiamo visto in precedenza, la maggior parte degli alloggi in quel territorio fanno registrare **prezzi di vendita inferiori ai 196.000$.** Sono **pochissimi** gli alloggi venduti al di **sopra di questa cifra.**
- **Situazione opposta**, invece, per il **College Creek**, zona del **Sud Edwards**, dove la maggioranza delle case fa registrare prezzi tra i **196.000$** ed i **243.000$** , con un buon gruppetto di case con prezzi **superiori** ai **243.000$** .
- Nonostante questo dato, **la zona del College Creek risulta la seconda per numero di case vendute.**

Tempo - Numero di case vendute per ogni anno.

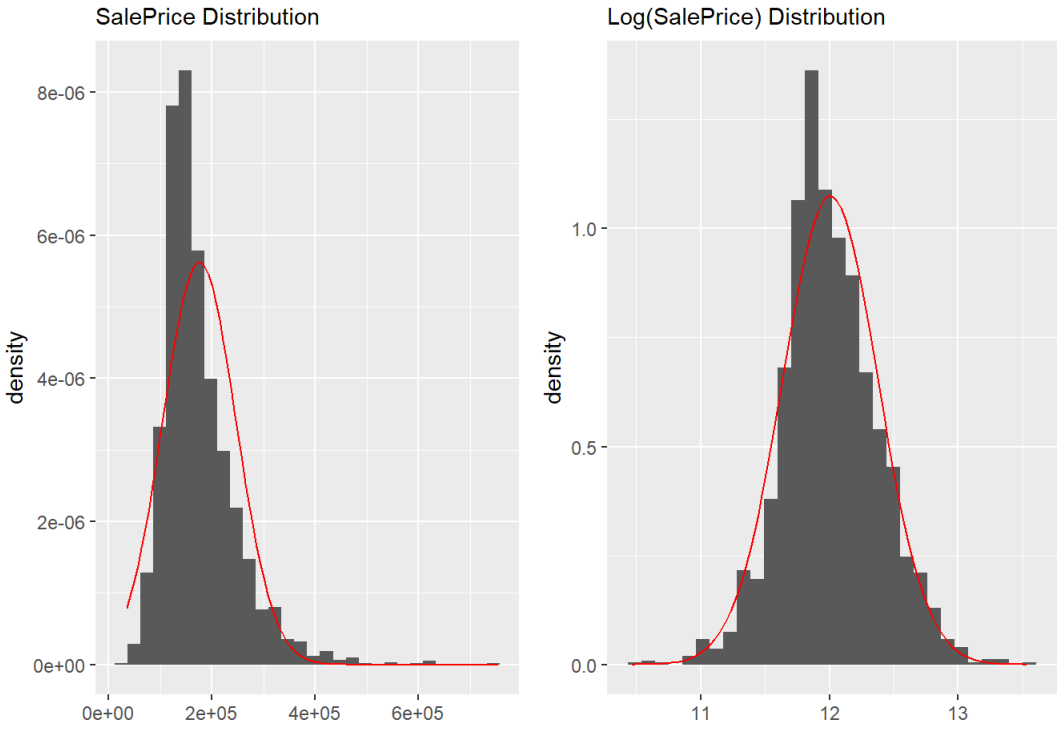


L'aspetto temporale incide sui prezzi?



- Sembra che il fattore temporale **non** abbia incidenza sul prezzo delle case.
- Interessante notare che tra il **2008 ed il 2010 non assistiamo ad un calo dei prezzi**, nonostante il campione rappresenti il periodo della **Grande Recessione, avvenuta tra il 2007 ed il 2013.**
- Infatti, questo avvenimento ebbe conseguenze dure per **l'intero Paese**, ma a quanto pare, **ad Ames i prezzi non hanno subito nessuna variazione e in positivo e in negativo, ma sono rimasti stabili.**
- Risulta altrettanto interessante che non solo i prezzi non sono diminuiti tra il 2008 ed il 2010, ma il **2009 risulta addirittura il miglior anno in termini di vendite.**
- Possiamo ipotizzare che il calo delle vendite nel **2010**, sia dovuto agli effetti della Grande Recessione oppure alla mancanza di osservazioni nel suddetto anno.

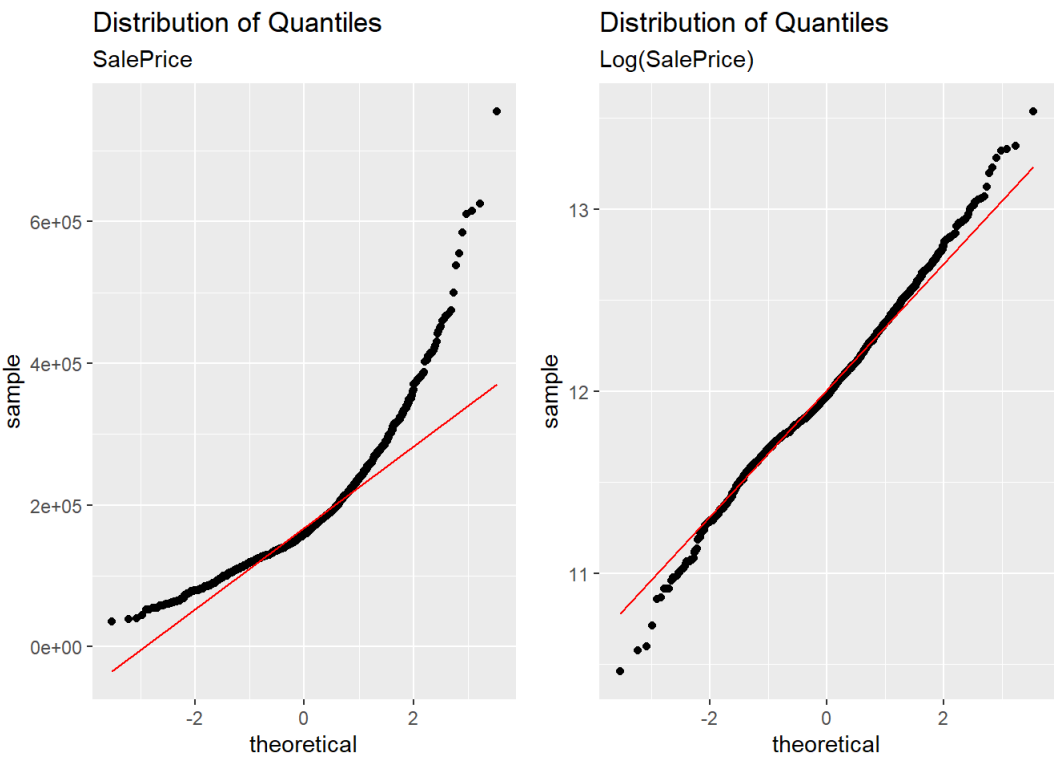
Analisi sul Prezzo e sul Log(Prezzo)



- Sembra avere maggiore senso considerare il **log dei prezzi**, che ci restituisce una **distribuzione quasi normalizzata dei dati.**

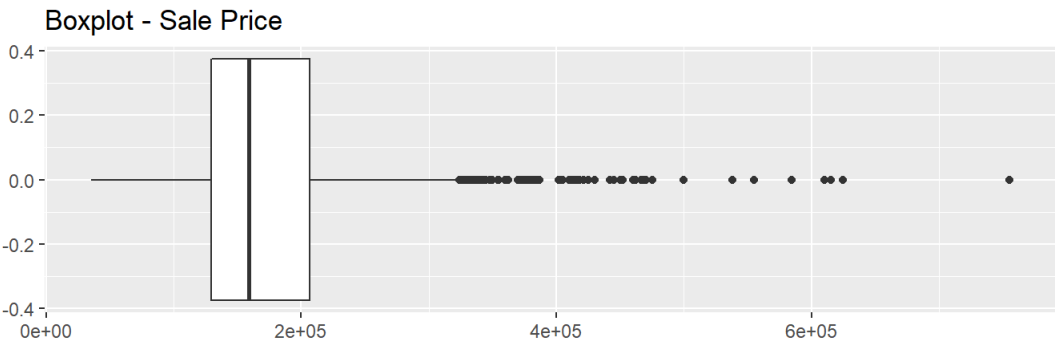
Verifichiamo ancora...

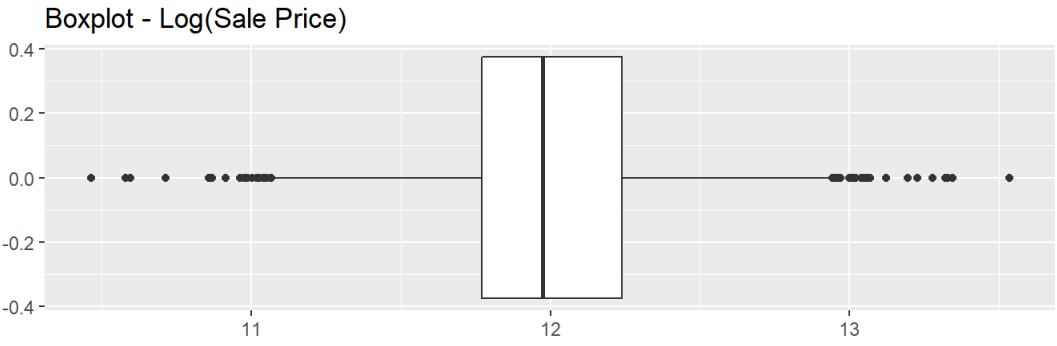
Q-Q Plot



- Anche da questa visualizzazione, notiamo che effettivamente la **trasformata logaritmica** rende la distribuzione dei quantili **quasi vicina a quella normale.**
- Si nota che la prima distribuzione **devia fortemente dalla normale**, specialmente nelle code.
- Si vede **un'asimmetria positiva**, infatti la **coda di destra** risulta molto **spessa rispetto alla normale**, proprio **come la coda di sinistra.**
- Il **comportamento** si avvicina a quello **normale**, solo nella **parte centrale** della distribuzione.

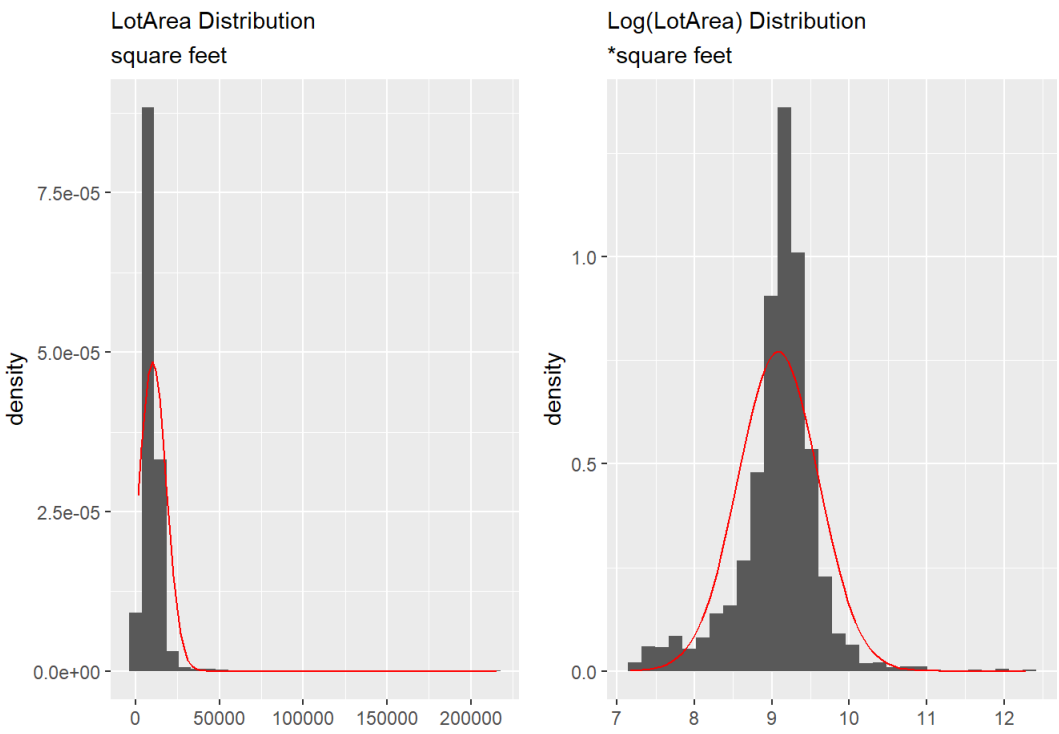
Boxplot





- Abbiamo la **conferma** di quanto detto in precedenza riguardo l'asimmetria positiva.
- Vediamo, infatti, che la distribuzione di **SalePrice presenta un'asimmetria positiva**, suggerita dal fatto che **il box di destra risulta maggiore in ampiezza rispetto al box di sinistra.**
- Diciamo anche che **questa ampiezza viene conservata anche con la trasformata logaritmica, ma risulta molto meno evidente.**

Distribuzione della dimensione del lotto



- **Aggiungiamo**, quindi, le **trasformate logaritmiche** al set di dati e le valuteremo poi per la costruzione del modello.

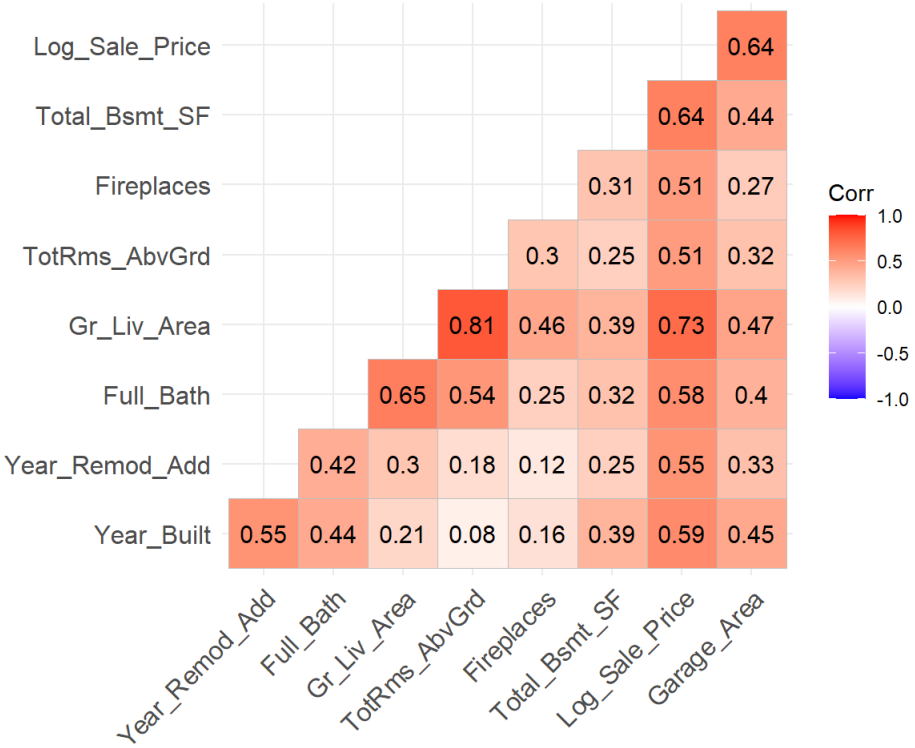
Quantili del Prezzo

```
##     SalePrice
## Q1     129500
## Q2     158750
## Q3     206925
## P95    312725
```

Dal calcolo dei quantili, invece, possiamo osservare che:

- Il **25% delle case meno costose**, ha un prezzo **inferiore a 130.000$** circa.
- Case che **superano i 312.000$ circa** , si trovano nell'ultimo 5% della distribuzione ed il **prezzo mediano** si aggira **sui 158.000$ circa.**

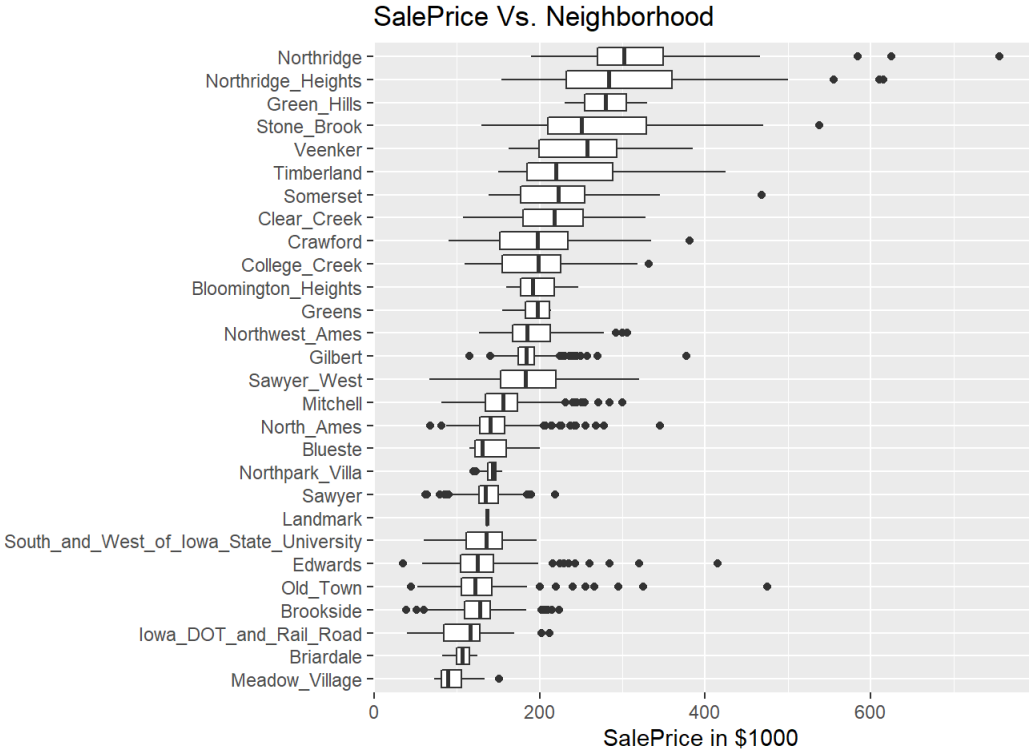
Quali caratteristiche hanno un'incidenza positiva sul Prezzo?



- Sul prezzo della casa, **incide** in maniera positiva la **dimensione**, in termini di superficie calpestabile, nel cui calcolo includiamo anche la **dimensione del garage e del seminterrato.**

- Hanno un ruolo importante anche il **numero di bagni** a disposizione, la **superficie interrata** e **l'anno di costruzione.**
- Caratteristiche come la presenza del **camino e l'anno di ristrutturazione, sono meno importanti** ma comunque da tenere in considerazione.

Prezzo per quartiere



- **Mediamente** il quartiere con il **prezzo maggiore alto risulta Northridge**, seguito dal quartiere **Veenker** e dal quartiere **Stone_Brook.**
- Il quartiere **meno costoso** risulta **Meadow_Village.**
- Infine, quartieri come **Crawford, Somerest e Stone_Brook,** presentano anche **un'importante variabilita dei prezzi**, a differenza di quartieri come **Gilbert o Mitchell** dove i prezzi sembrano essere meno variabili

---

Step 3 - Features Engineering

Per questo step, prendiamo delle variabili che si riferiscono alle stesse caratteristiche per **combinarle tra loro ed ottenere un'informazione sintetica e facile da interpretare.**

- Alcune trasformazioni riguardano le seguenti variabili:
  - **TotalPorch** = Open_Porch_SF + Screen_Porch + Enclosed_Porch + Three_season_porch.
  - **TotalBath** = Bsmt_Full_Bath + Bsmt_Half_Bath + Full_Bath + Half_Bath.
  - **TotalFLRSF** = First_Flr_SF + Second_Flr_SF.
- Con queste trasformazioni, abbiamo raccolto in **TotalPorch** tutta **l'informazione relativa all'area del porticato**, indipendentemente se l'area risulta recintata, aperta o altro.
- Si sintetizza con **TotalBath**, invece, l'informazione sul **numero di bagni indipendentemente dalla collocazione e dalla finitura.**
- Infine, con **TotalFLRSF**, abbiamo l'informazione sui metri **quadrati del primo e del secondo piano.**
- Altre trasformazioni riguardano:
  - **TotalOverall** = Overall_Cond_New * Overall_Qual_New.
  - **TotalKitchen** = Kitchen_Qual_New * Kitchen_AbvGr.
  - **TotalGarage** = Garage_Cond_New * Garage_Qual_New.
  - **TotalExter** = Exter_Cond_New * Exter_Qual_New.
- **TotalOverall** indica l'aspetto **qualitativo della casa** in base ai **materiali, ed alla condizione della casa.**
- **TotalKitchen** e **TotalGarage** indicano le **condizioni** rispettivamente della **cucina** e del **garage.**
- Infine, **TotalExter** contiene l'informazione sull'aspetto **qualitativo dei materiali dell'esterno degli alloggi.**

Step 4 - Models

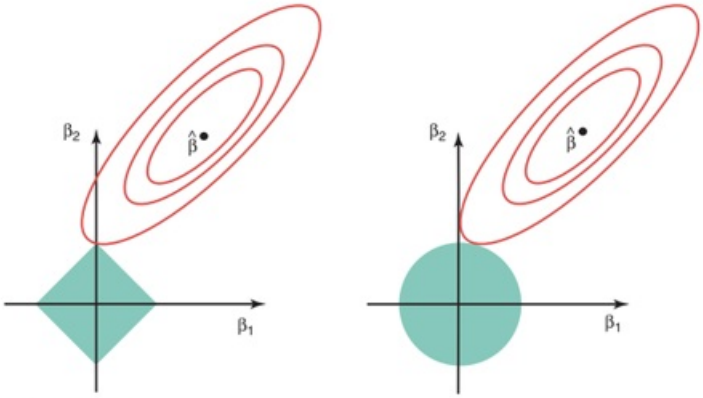
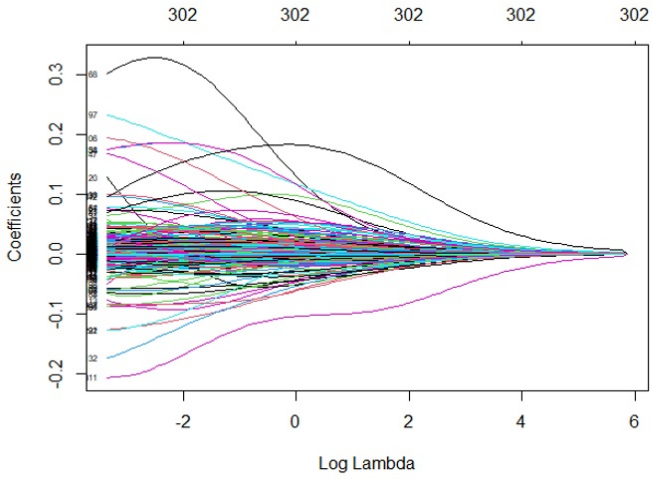


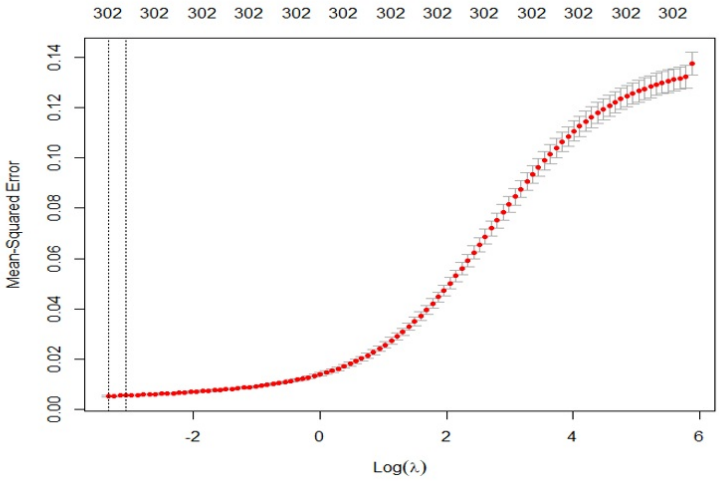
FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

- Sia la stima ottenuta con il **metodo Ridge** che con il **metodo Lasso** sono **regressioni penalizzate**; ovvero **metodi che si basano sulla penalizzazione della funzione obiettivo**, aggiungendo a quest'ultima un **termine di penalizzazione** che agisce sui parametri di Beta.
- Il problema di ottimizzazione, quindi, diventa un **problema di ottimizzazione vincolata**, dove il vincolo viene rappresentato dalla **somma dei quadrati dei coefficienti, ponendola minore di una certa costance C** con la Ridge Regression.
- Utilizzando il **Lasso**, invece, il **vincolo** ha una forma valore **assoluto** e non quadratico. Questo si traduce **geometricamente** in un **rombo** e non in una circonferenza.
- Il vantaggio di utilizzare una regressione di tipo **Lasso**, e non una regressione Ridge, sta nel fatto che con il Lasso **i coefficienti non vengono solo tirati verso lo 0, ma vengono effettivamente posti a zero.** Questa differenza fa emergere un ulteriore aspetto discriminante tra i due approcci; ovvero che l'approccio **Lasso ci permette di ottenere contemporaneamente sia le stime, sia una selezione delle variabili.**
- Inoltre, **anche per i coefficienti che sono diversi da zero**, con la regressione **Lasso vengono comunque tirati maggiormente verso lo zero**, rispetto a quanto avviene con la regressione Ridge.

Ridge

Stime Ridge





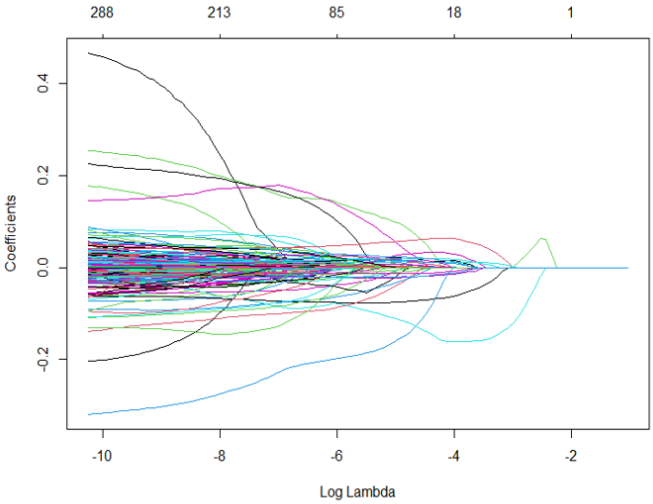
Lambda ottimale utilizzando la Cross-Validation

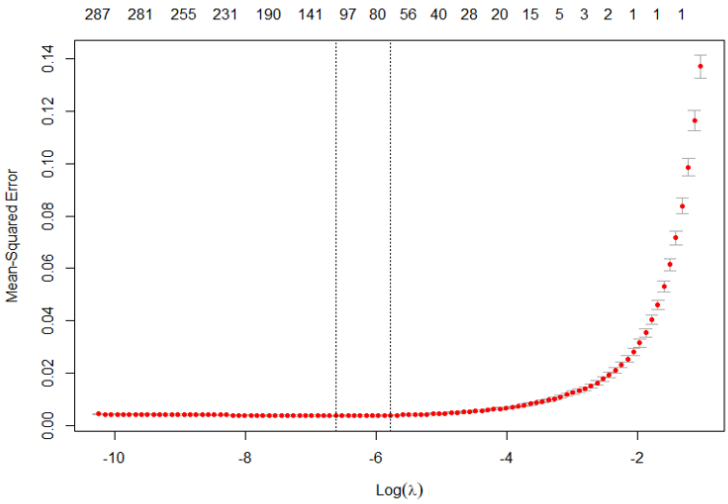
Valore minimo di Lambda

```
## [1] 0.03546255
```

Lasso

Stime Lasso





Lambda ottimale utilizzando la Cross-Validation

Valore minimo di Lambda

```
## [1] 0.001335142
```

Performance con Lambda.min

```
##       RMSE       R2
## 1 0.05421903 0.9791042
```

Elastic Net

Stime Elastic Net

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \text{ s.t. } (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t$$

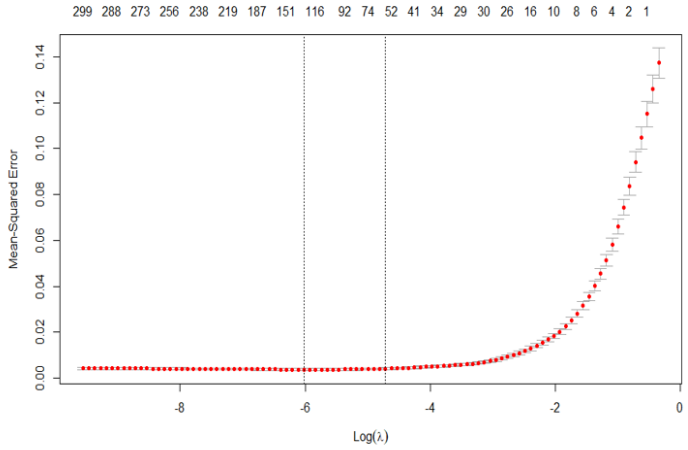
where $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$

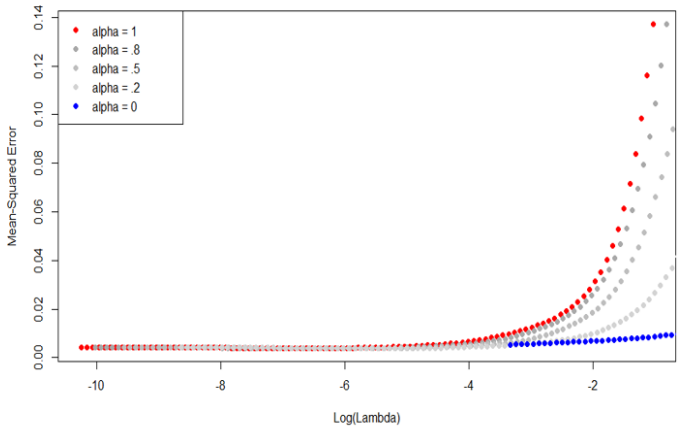
- Con l'**Elastic Net**,si vuole prendere il meglio dell'approccio Ridge e del Lasso, quindi trovare un **compromesso tra la penalizzazione in norma L1 e quella in norma L2.**
- Nella penalizzazione, viene aggiunto un **nuovo parametro**, ovvero **Alpha**, il cui valore ci permette di **tirare la penalizzazione stessa verso un vincolo di tipo ridge o di tipo lasso.**
- Il **vantaggio** di quest'approccio sta nel fatto che possiamo sia fare **variable selection**, sia superare i problemi dell'approccio Lasso.
- Infine, con questo approccio **non abbiamo limitazioni sul numero di variabili selezionate**(differenza con l'approccio lasso) ed in presenza di variabili con correlazioni di gruppo, non abbiamo problemi.
- Questo vantaggio lo paghiamo in **termini di scelta**, in quanto i parametri da fissare saranno due: **Lambda** per il livello di shrinkage ed **Alpha** per l'elasticity della net.
- Per la **scelta di Lamda**, utilizziamo la **K-Fold Cross Validation.** Questa tecnica consiste nella **suddivisione dell'insieme di dati totale in k parti** di uguale dimensione e, ad ogni passo, la **k-esima parte dell'insieme di dati viene ad essere quella di convalida**, mentre la **restante parte costituisce sempre l'insieme di addestramento**. In questo modo, si allena il modello per ognuna delle k parti, evitando quindi problemi di sovradattamento, ma anche di campionamento asimmetrico del campione osservato, che succede tipicamente quando si suddividono i dati in due sole parti.
- Ricordiamo che i dati vengono casualmente partizionati in k-folds(sottocampioni) che non condivideranno alcuna osservazione:

$$\text{Fold-1} \cap \text{Fold-2} \cap \ldots \cap \text{Fold-k} = \emptyset$$

- La grandezza del **train** e del **test** viene determinata da **k.** Infatti in ogni fold avremo una frazione di n/k dati, quindi: avremo **1 - n/k** per la frazione di dati assegnata al training set e **n/k** per la frazione di dati assegnata al test set.
- **Quale errore stima CV?**
- CV stima l'**Err**, ovvero **l'Expected Test Error**, inteso anche come il **valore atteso del test error rispetto a tutti i possibili train sample.**

Lambda ottimale utilizzando la Cross-Validation





Performance con Lambda.min

```
##              RMSE       R2
## el 1   0.05421903 0.9791042
## el 0.8 0.05448815 0.9789053
## el 0.5 0.05369061 0.9795199
## el 0.2 0.05294828 0.9800970
## el 0   0.05991732 0.9746377
```