# USA Clima

EMANUELE DI PIETRO

L’analisi svolta sui dati ha come scopo di vedere l’esistenza di relazioni tra la posizione geografica di un campione di città degli Stati Uniti e alcune variabili. Il dataset è composto da 60 osservazioni e 6 variabili, tutte quantitative. I dati analizzati sono stati presi in considerazione dalla libreria di dati online D.A.S.L. (Data And Story Library):

*https://dasl.datadescription.com/datafile/city-climate/*

Le variabili contenute nel dataset sono:

-Latitudine: distanza angolare del punto dall’equatore (°)

-Longitudine: distanza angolare del punto dal Meridiano di Greenwich (°)

-JanTF: Temperatura media di gennaio (°F)

-JulyTF: Temperatura media di luglio (°F)

-RelHum: Umidità relativa (%)

-Rain: Quantità di pioggia (l/m2)

Inoltre, oltre a queste variabili quantitative è stata creata una variabile qualitativa State, a 4 livelli, riguardo alcune macroregioni degli USA:

-Stati medio-occidentali

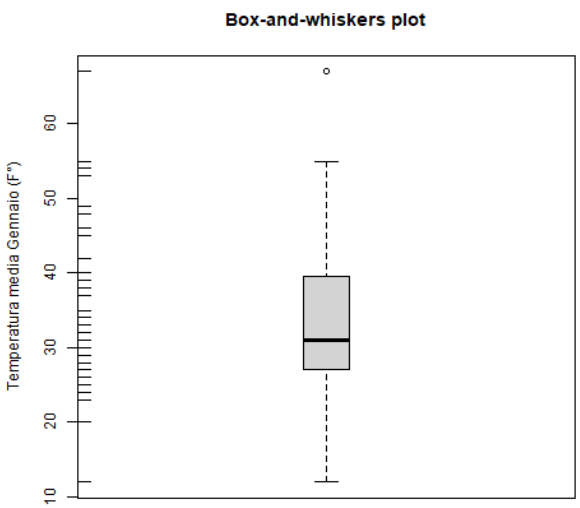
-Stati occidentali

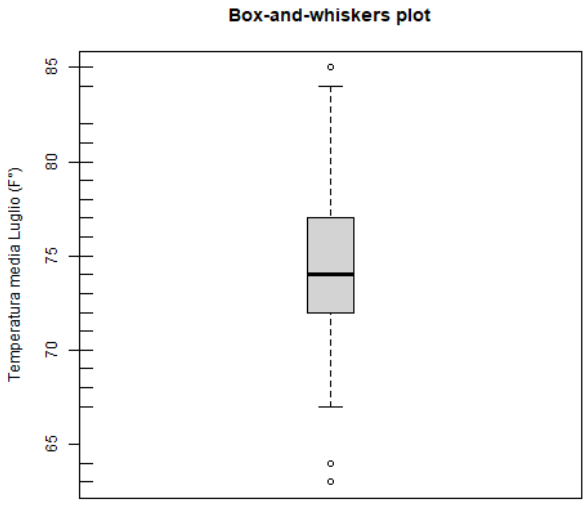
-Stati nord-orientali

-Stati del sud

# Analisi esplorativa univariata

Partendo da un'analisi esplorativa delle singole variabili è possibile creare dei 'boxplot' per le variabili quantitative e un 'diagramma a nastri' per quella qualitativa. Avremo:

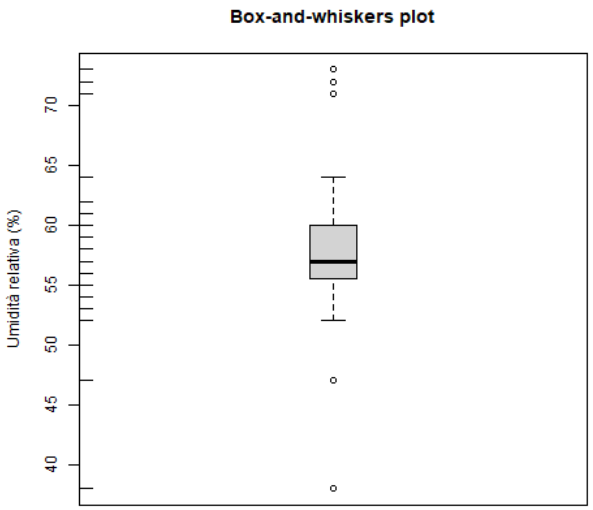


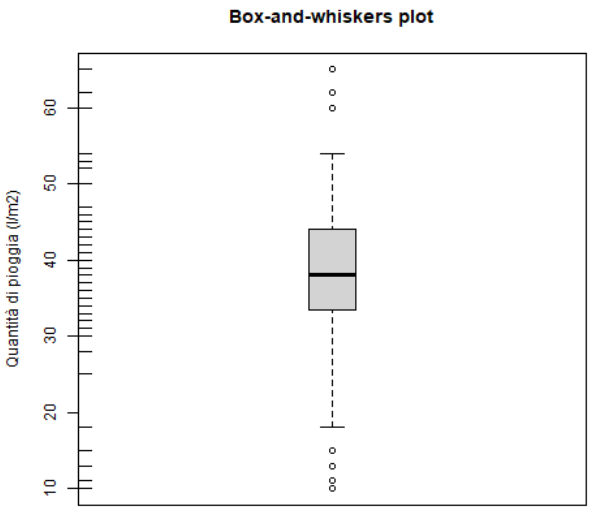




Dal 1° boxplot si può intravedere una leggera asimmetria positiva, con mediana pari a 31 e media 33.8; salta all'occhio soprattutto il valore anomalo (67°F) relativo alla città di Miami, ma ciò ha senso poiché è la città più a sud-est del campione, quindi più a ridosso del Tropico del Cancro. Il valore più basso lo raggiunge la città di Minneapolis (Minnesota) con 12°F, città al nord degli USA, al confine con il Canada, più vicina all'Artide. Il 50% delle osservazioni è compreso tra 27° e 39.5°. Per quanto riguarda il 2° boxplot si ha una mediana di 74, cosi come la media 74.41, valori molto simili e infatti sembra esserci anche una certa simmetria; il 50% qui è compreso in un range molto più piccolo, ossia 72°-77°, questo ci fa capire che per quanto riguarda il periodo estivo c'è più omogeneità, con temperature non troppo differenti in tutto il paese

(nonostante la vastità), al contrario del periodo invernale. I valori anomali qui sono 3: Dallas (85°) città del Texas, al sud del paese e Seattle e San Francisco, due città situate sulla costa occidentale dove il clima risulta più fresco e le temperature non raggiungono alte quote in estate.

Per le altre 2 variabili quantitative abbiamo:




Nel 3° boxplot, relativo all'umidità abbiamo che la mediana e la media sono quasi uguali, rispettivamente 57 e 57.75, anche qui sembra esserci una certa simmetria e le osservazioni sono quasi tutte comprese tra 52% e 64%, tranne 5 valori anomali, il più basso è Denver (38%), città del Colorado, nel bel mezzo delle montagne rocciose, mentre i valori anomali più elevati sono Portland, Seattle, San Francisco, tutte situate sulla costa occidentale, dove quindi l'umidità sembra essere più elevata.

L'ultimo boxplot, relativo ai mm mensili medi si nota una leggera asimmetria negativa, però mediana e media non differiscono più di tanto, rispettivamente 38 e 38.51; il range qui è molto elevato quindi possiamo ipotizzare che non piove in modo omogeneo in tutto il territorio, infatti

abbiamo alcuni valori anomali molto alti (York, Worcester e Willington, tutte sulla costa orientale) e anche molto bassi (San Diego, San Jose, Los Angeles, città costa occidentale e Denver, montagne rocciose).

```
> variance(JanTF)
[1] 101.3146
> skewness(JanTF)
[1] 0.9910352
> kurtosis(JanTF)
[1] 4.043077
```

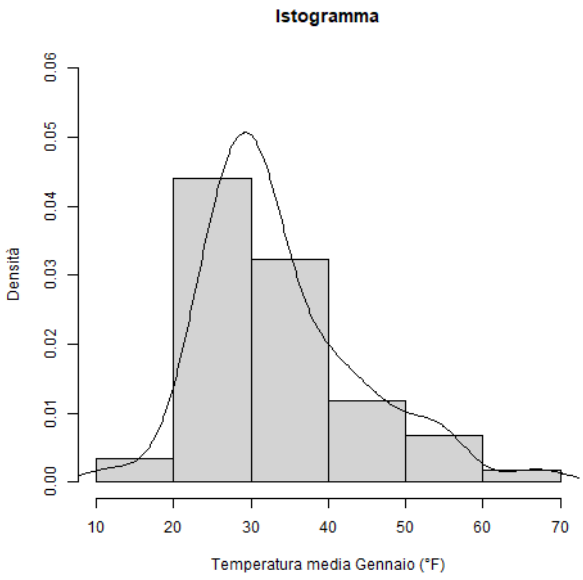
```
> variance(JulyTF)
[1] 20.81758
> skewness(JulyTF)
[1] 0.06480433
> kurtosis(JulyTF)
[1] 2.941307
```

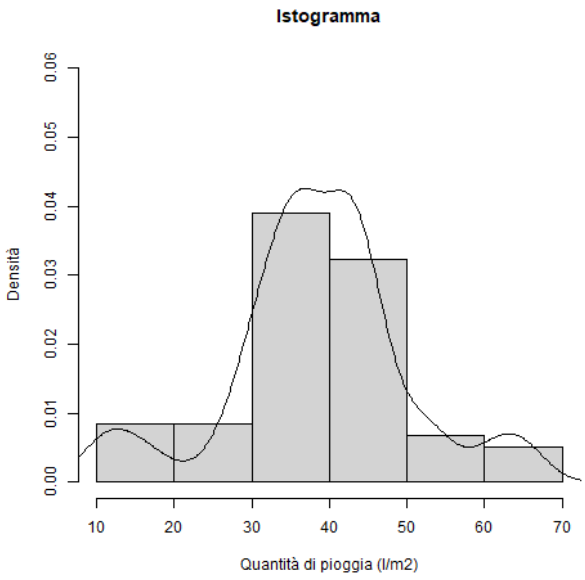
```
> variance(RelHum)
[1] 28.46079
> skewness(RelHum)
[1] 0.2010112
> kurtosis(RelHum)
[1] 6.883367
```

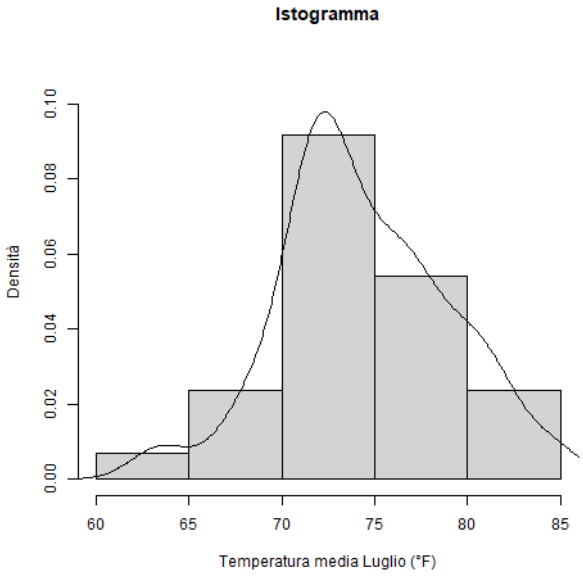
```
> variance(Rain)
[1] 131.6737
> skewness(Rain)
[1] -0.1769058
> kurtosis(Rain)
[1] 3.802736
```

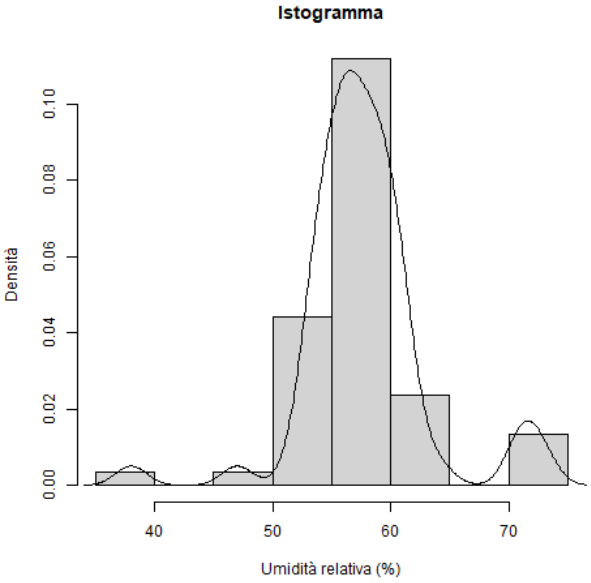
Dalla stima degli indici della ‘varianza’, ‘asimmetria’ e ‘curtosi’ si può vedere quali sono le distribuzioni che tendono più ad una normale, ovvero quelle con asimmetria 0 e curtosi 3, che sono la variabile relativa alle temperature medie di luglio e, in modo meno evidente, quella relativa alla quantità media di pioggia mensile; quest’ultima ha una leggera asimmetria negativa che si intravedeva già dal boxplot, così come la leggera asimmetria positiva nel 1° boxplot.

Di queste variabili quantitative è possibile vedere anche gli istogrammi e le relative funzioni di densità:

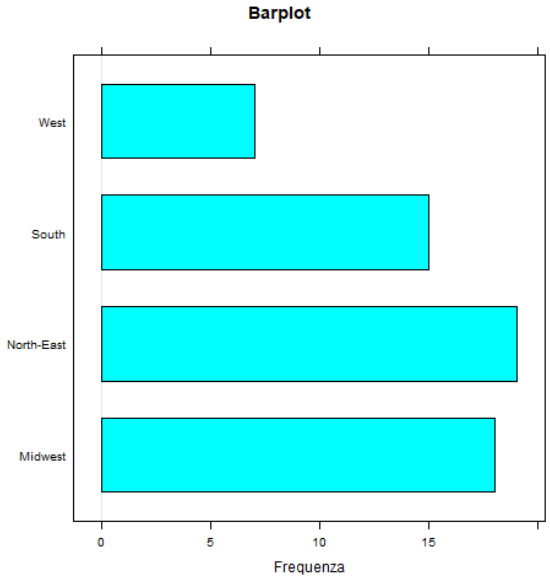





Dalla prima funzione di densità si nota molte bene l'asimmetria positiva, mentre graficamente non sembra molto evidente la normalità della variabile JulyTF.

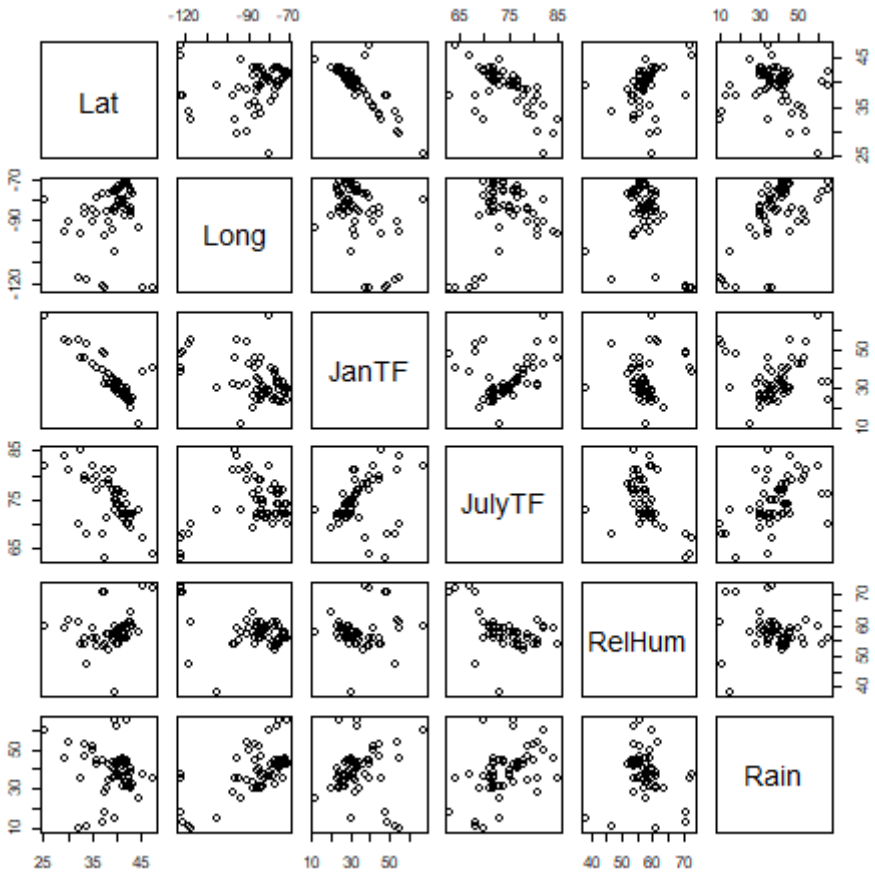



```
> table(State)
State
   Midwest North-East      South       West
        18         19         15          7
```

Il 'barplot', per quanto riguarda la variabile qualitativa, ci dice (graficamente) solamente quante osservazioni ci sono per ogni livello, così come distribuzione di frequenza.

# Analisi bivariata e multivariata

Passando ad un'analisi bivariata si può iniziare a vedere se c'è (o meno) qualche dipendenza lineare tra le variabili; per prima cosa facciamo una 'matrice degli scatter plot':



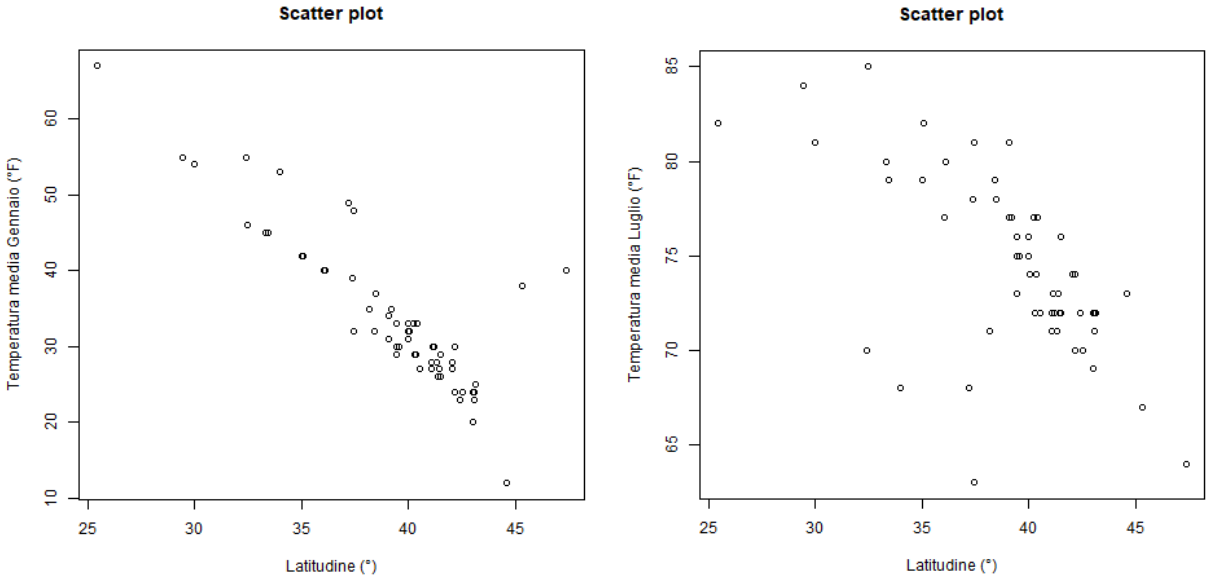

Dalla matrice degli scatter plot possiamo notare la buona dipendenza lineare tra JanTF e Lat, inoltre sembrano esserci dipendenze lineari, ma meno evidenti, anche tra JanTF e JulyTF e tra JulyTF e Lat.

Volendo quantificare queste relazioni è possibile fare una matrice di ‘correlazione:

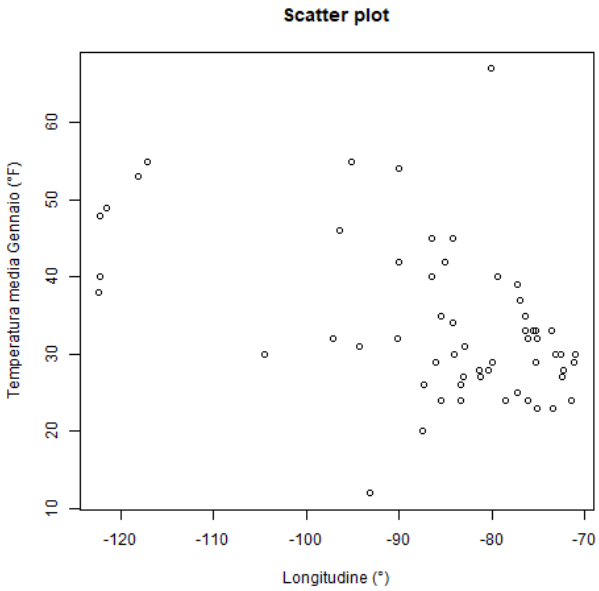
```
> cor(Data[, 2:7])
              Lat       Long       JanTF     JulyTF      RelHum        Rain
Lat     1.0000000  0.1845849 -0.85731347 -0.6225834  0.21224529 -0.16060495
Long    0.1845849  1.0000000 -0.47585997  0.2986981 -0.39626765  0.61909404
JanTF  -0.8573135 -0.4758600  1.00000000  0.3221455  0.08552171  0.05856608
JulyTF -0.6225834  0.2986981  0.32214555  1.0000000 -0.44139661  0.47225673
RelHum  0.2122453 -0.3962677  0.08552171 -0.4413966  1.00000000 -0.11777277
Rain   -0.1606050  0.6190940  0.05856608  0.4722567 -0.11777277  1.00000000
```

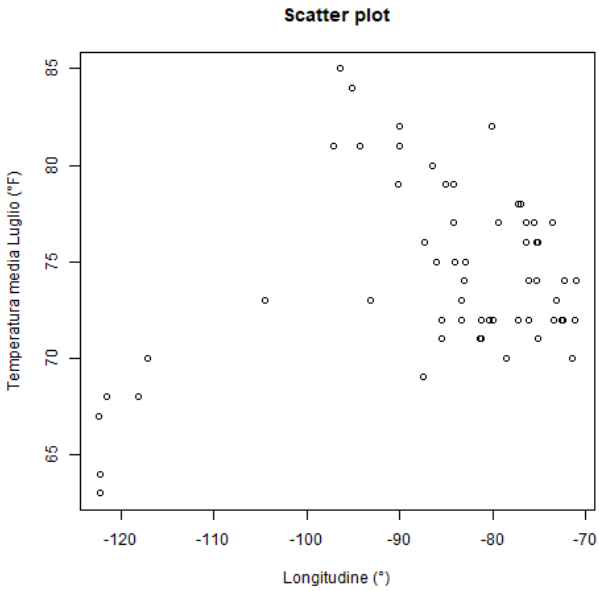
Tale matrice conferma quanto sospettato in precedenza e, inoltre, mostra anche una buona dipendenza lineare tra Long e Rain, tra JulyTF e Rain ed infine JulyTF e RelHum; la relazione tra JanTF e JulyTF, invece, non è poi così forte come sembrava.

Possiamo quindi andare a vedere gli scatter plot singolarmente, focalizzandoci sulle relazioni appena citate:



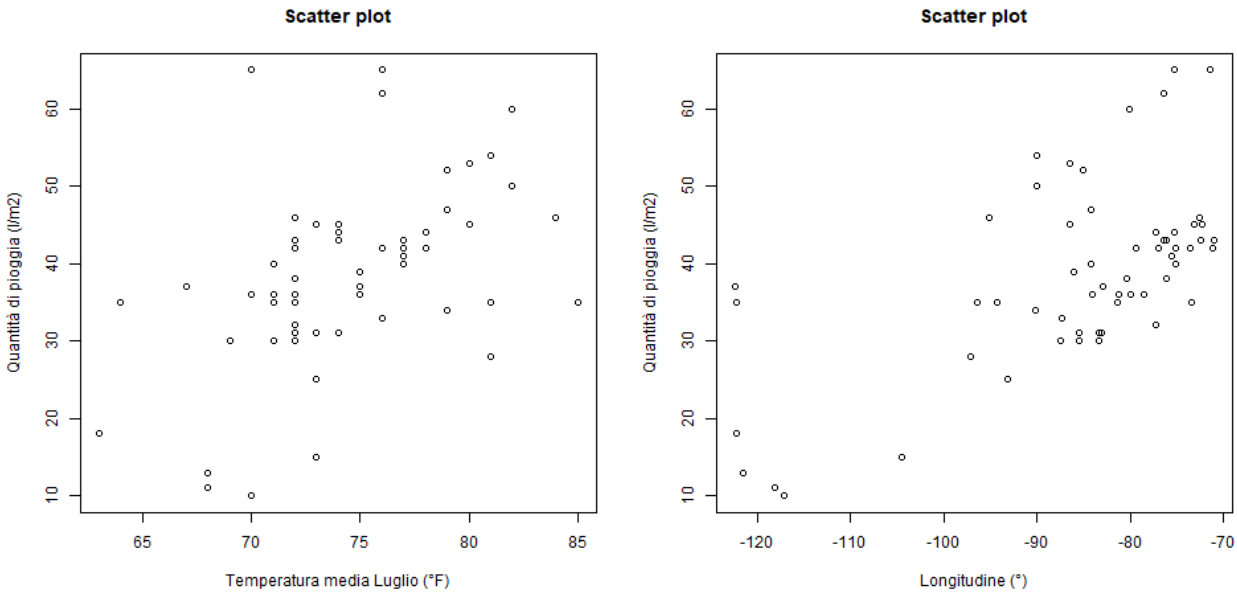

All’aumentare della ‘Latitudine’, quindi spostandoci da sud a nord, si nota che le temperature medie di gennaio diminuiscono e ciò è molto ragionevole poiché ci si allontana sempre più dal Tropico del Cancro, avvicinandosi al Canada e quindi al polo; stessa cosa per le temperature media di luglio, anche se qui c’è un po' più di dispersione.



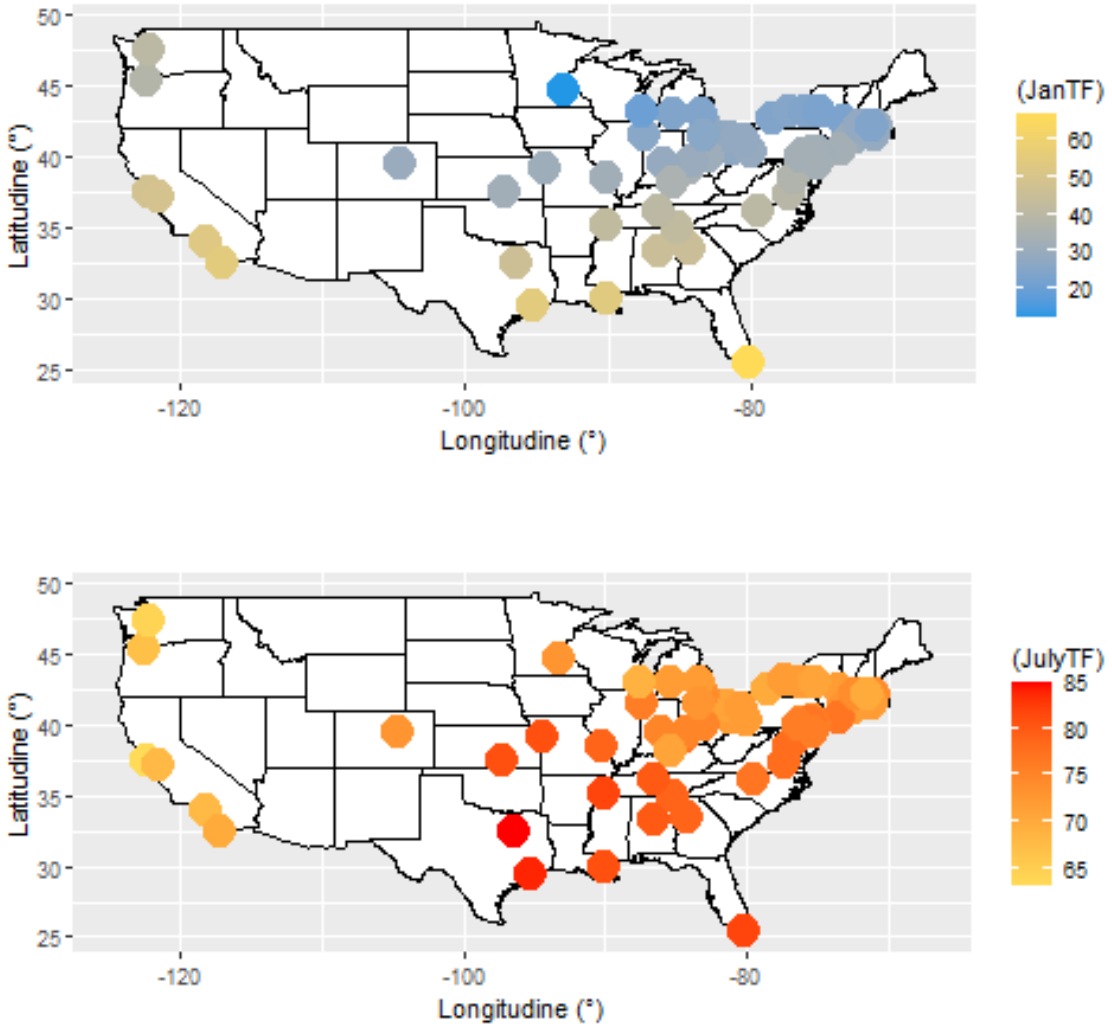

Con la ‘longitudine’ la situazione è un po' diversa, un po' meno marcata, però anche qui con valori più alti, quindi spostandoci da sinistra a destra, le temperature medie di gennaio tendono a diminuire (tranne nel caso di Miami che raggiunge 67°F nonostante sia molto a est, questo perché è anche molto a sud); per quanto riguarda le temperature medie di luglio la situazione è simile poiché al crescere della ‘longitudine’ le temperature diminuiscono, tranne che nel caso delle città situate più a occidente (-120° circa), infatti lì le estati non raggiungono temperature altissime.

Infine, analizziamo altri 2 scatter plot molto interessanti:




il primo mostra una dipendenza positiva mediocre, all'aumentare delle temperature estive aumentano anche i mm di pioggia caduti, mentre nel secondo caso i mm aumentano allo spostarsi da ovest a est (caso particolare di Portland e Seattle che nonostante siano a ridosso del 120° sono abbastanza piovose con rispettivamente 37 e 35mm, città che comunque sono molto a nord, a ridosso del Canada).

Per visualizzare al meglio la situazione delle temperature (sia di gennaio che di luglio) lungo il territorio è conveniente sviluppare una mappa sulla base della latitudine e della longitudine il cui relativo scatter plot ci dà la posizione geografica delle unità statistiche, ossia le città.

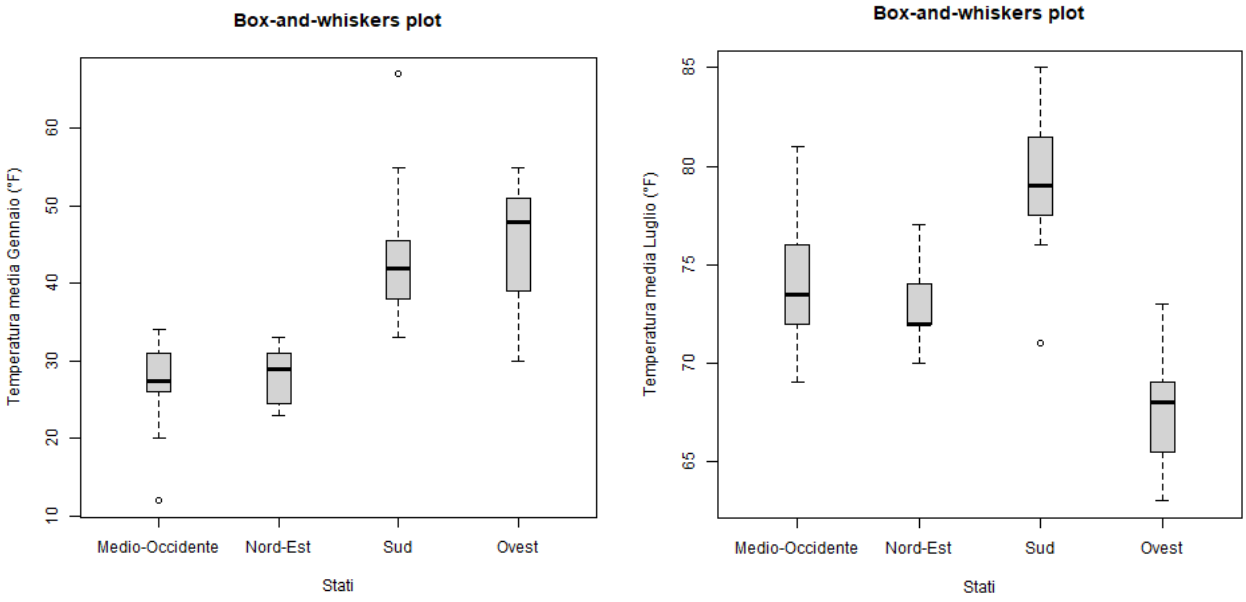



La mappa riassume molto bene quanto detto finora, partendo dalla situazione di gennaio vediamo che scendendo al sud le temperature aumentano (Miami in Florida raggiunge addirittura le minime estive) e di conseguenza avvicinandosi verso il Canada diminuiscono; sulla costa occidentale, inoltre, le temperature sembrano essere in generale più elevate rispetto al versante orientale, salendo con il salire della latitudine. La situazione in estate mostra invece delle temperature molto elevate al sud-est, in particolar modo in Texas (Dallas e Houston) e in Florida; nella parte nord-est le temperature sembrerebbero molto omogenee tra gli stati,

mentre sul versante occidentale si hanno le temperature più basse, dove quindi le temperature non variano di molto nell'arco dell'anno.

# Analisi condizionata

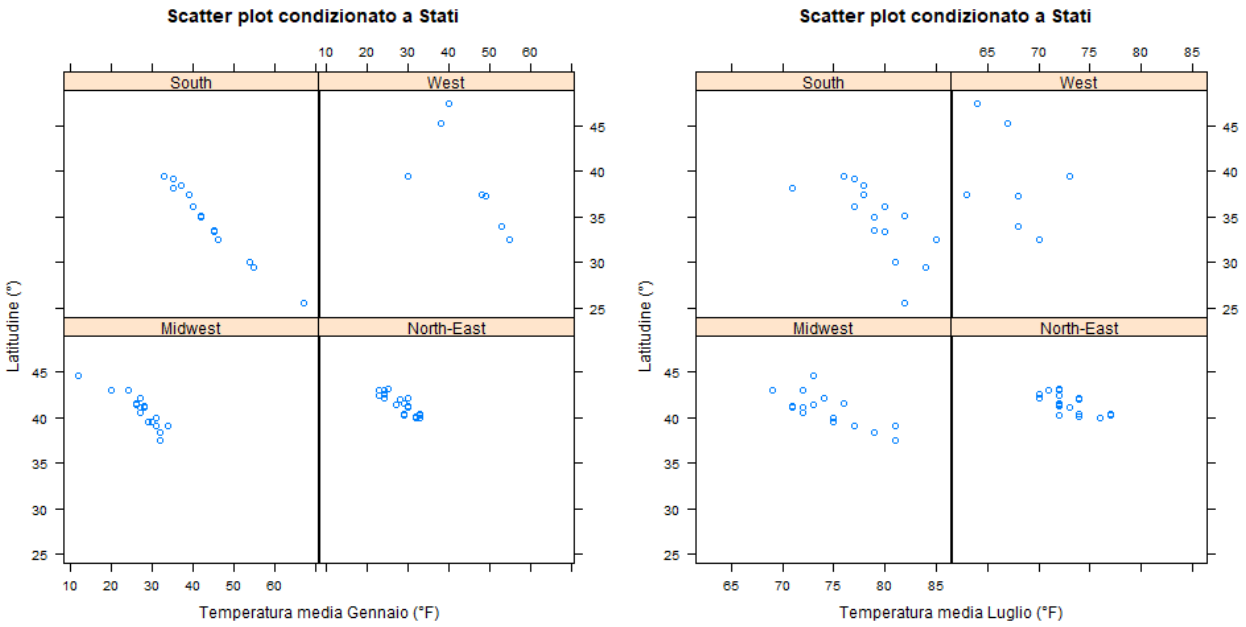
Adesso passiamo a vedere il comportamento delle variabili quantitative condizionate alla variabile qualitativa State. Partiamo sempre con dei boxplot:



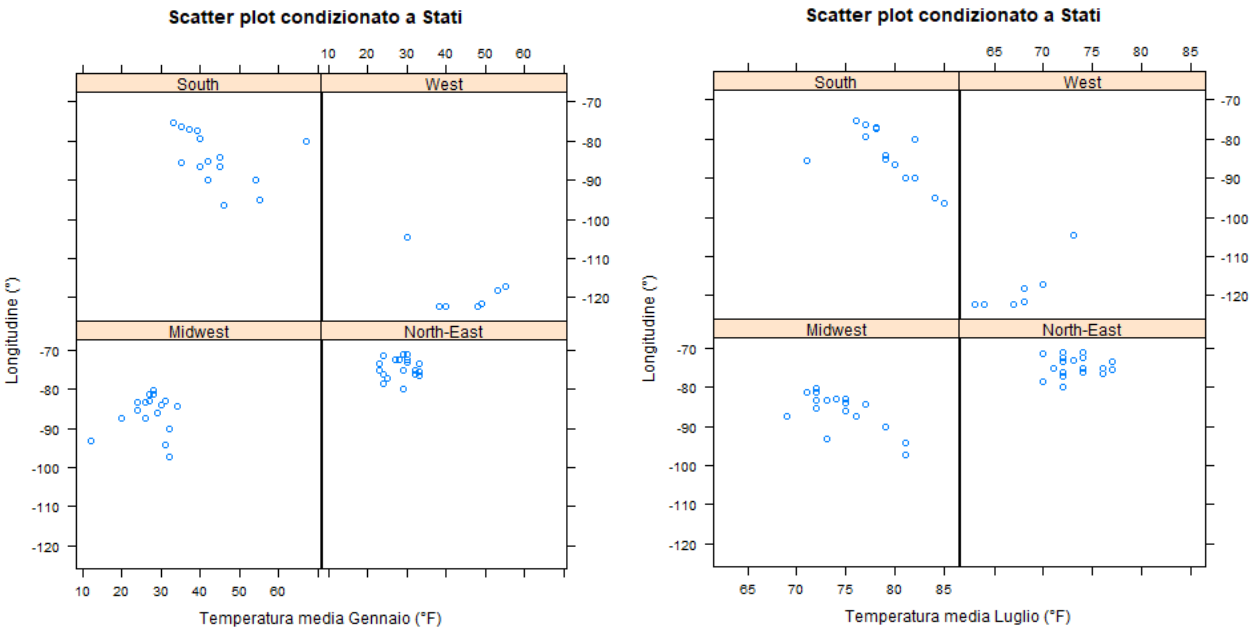

Questi boxplot non fanno altro che rafforzare quanto detto prima con le mappe geografiche, infatti le temperature più basse in gennaio si hanno nel Medio-Occidente ('Midwest') dove si ha il valore anomalo Minneapolis e le temperature più alte si hanno al sud (valore anomalo Miami) e ad occidente. Per il periodo estivo invece abbiamo che ad ovest le temperature non si alzano molto ma restano molto simili a quelle invernali, mentre nelle restanti parti del paese si alzano e anche di molto. Il valore

anomalo è Louisville, città praticamente situata a metà strada, quindi la si potrebbe considerare come la città del sud meno meridionale.

Accorpando più variabili insieme è possibile fare degli scatter plot condizionati.




Dal primo scatter plot risultano relazioni negative molto forti nel Sud, Medio-Occidente e Nord-Est, mentre nel secondo sembra esserci una relazione molto buona nel Medio-Occidente. Nel Nord-Est le temperature estive sono molto omogenee; anche nel Sud probabilmente c'è una mediocre dipendenza negativa, mentre a Occidente c'è molta dispersione (si hanno anche poche osservazioni per questa modalità).

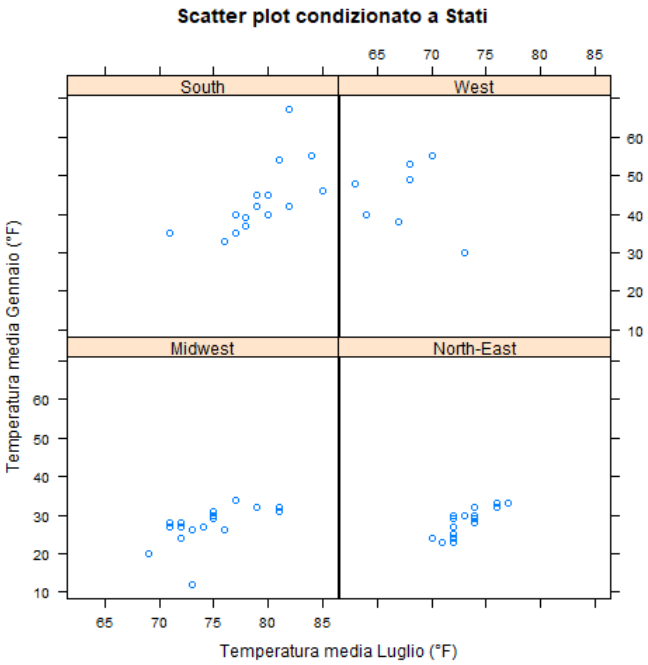



In questi altri due scatter plot (questa volta temperature medie di gennaio e luglio relazionate alla longitudine e condizionate a State) abbiamo situazioni abbastanza diverse:

-1° scatter plot: non si hanno molte dipendenze lineari, forse una piccola dipendenza c'è al Sud. Nel Nord-Est la temperatura di gennaio diventa quasi una costante.

-2° scatter plot: qui le dipendenze sono già più evidenti, soprattutto al Sud e nel Medio-Occidente; curiosa è però la situazione dell'Occidente, poiché a differenza delle altre parti qui c'è una relazione lineare positiva, ossia dove all'aumentare della longitudine aumenta la temperatura. Nel Nord-Est anche qui le temperature sono abbastanza costanti.

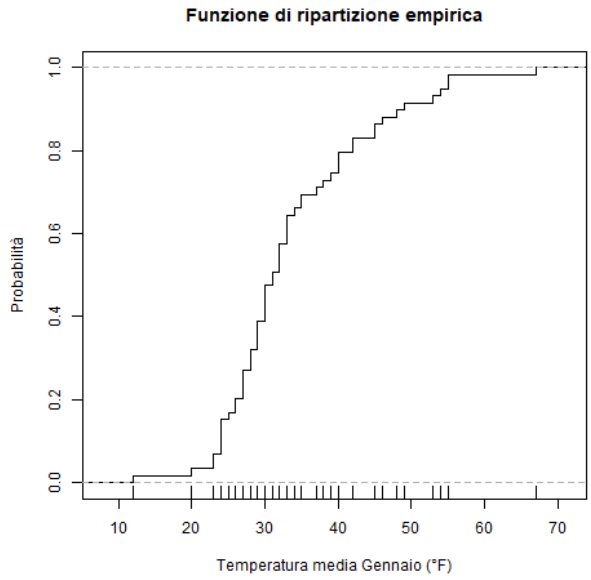
Infine, consideriamo 2 variabili che prima non ci hanno dato troppa dipendenza, ossia JanTF e JulyTF:

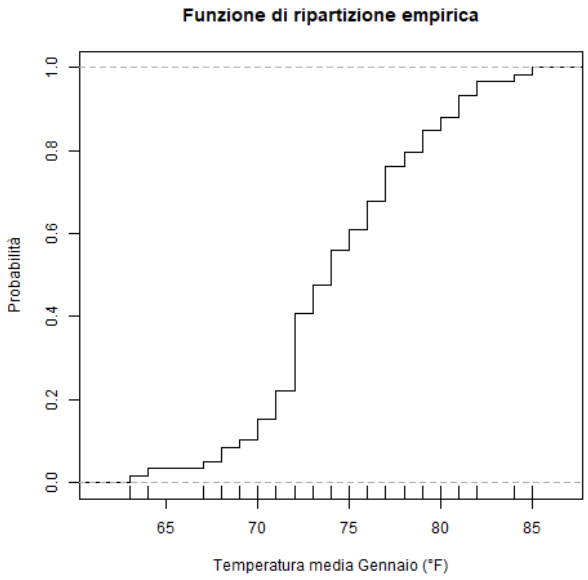


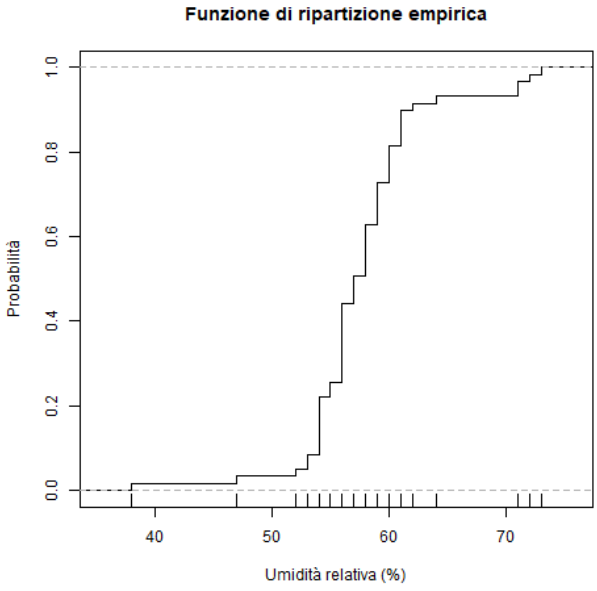

Dalla ‘matrice di correlazione’ vista in precedenza avevamo visto che la correlazione tra queste 2 variabili era pari a 0.32, quindi non molto forte, però guardando bene lo scatter plot sembra che questa dipendenza sia più marcata, infatti nel Medio-Occidente, nel Nord-Est e anche nel Sud sembra esserci una relazione lineare positiva buona, mentre ad Occidente non sembra esserci nessuna dipendenza lineare, c’è molta dispersione e proprio questa dispersione potrebbe aver influito in modo negativo sul ‘coefficiente di correlazione’.

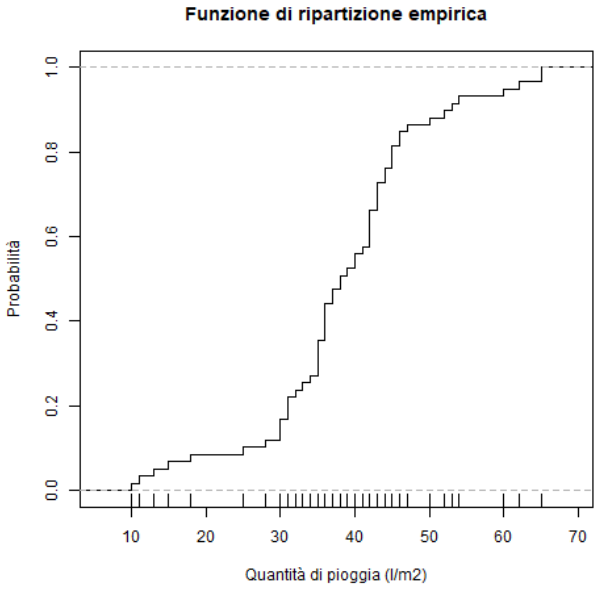
# Funzione di ripartizione empirica

La funzione di ripartizione empirica è una statistica “distribution-free” che fornisce la percentuale di osservazioni campionarie minori o uguali ad un determinato valore. I grafici forniti da R per le variabili quantitative sono:



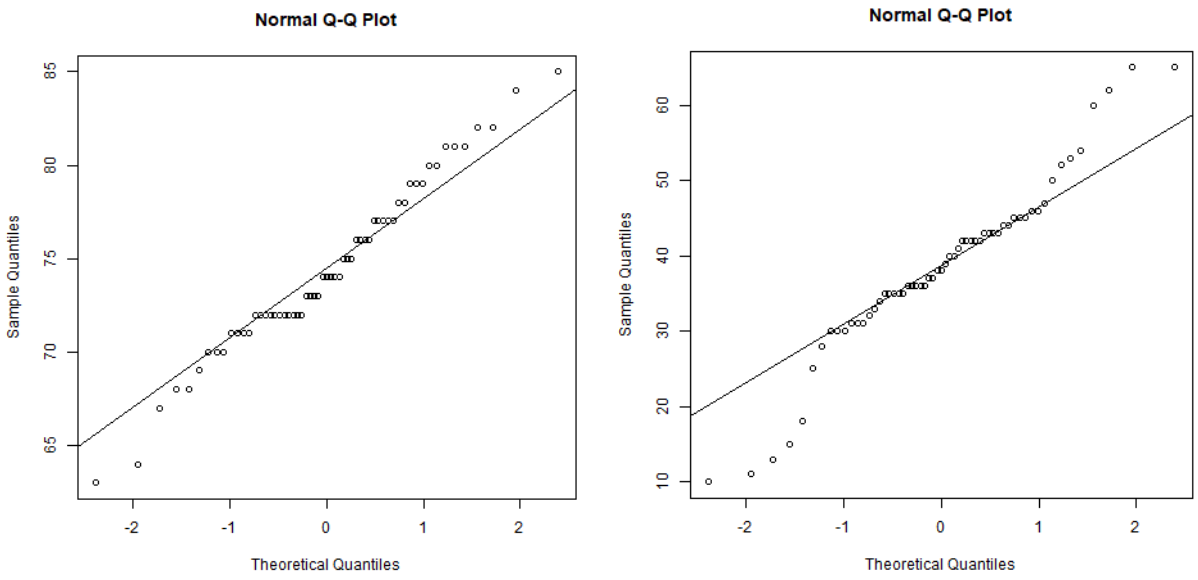



# Massima verosimiglianza

Dagli istogrammi, dalla stima degli indici di 'asimmetria' e di 'curtosi' e dalla funzione di ripartizione sembrava ancora più evidente la normalità di alcune distribuzioni, in particolar modo quella della variabile JulyTF e in modo meno marcato la variabile Rain; assumendo un campionamento casuale da una variabile casuale normale, le stime di massima verosimiglianza dei parametri μ e $\sigma^2$ risultano rispettivamente la media e la varianza stimate sul campione, , ossia $\bar{x}$ e $s^2$:

```
> mean(JulyTF)  > mean(Rain)
[1] 74.40678    [1] 38.50847
> var(JulyTF)   > var(Rain)
[1] 21.1765     [1] 133.9439
```

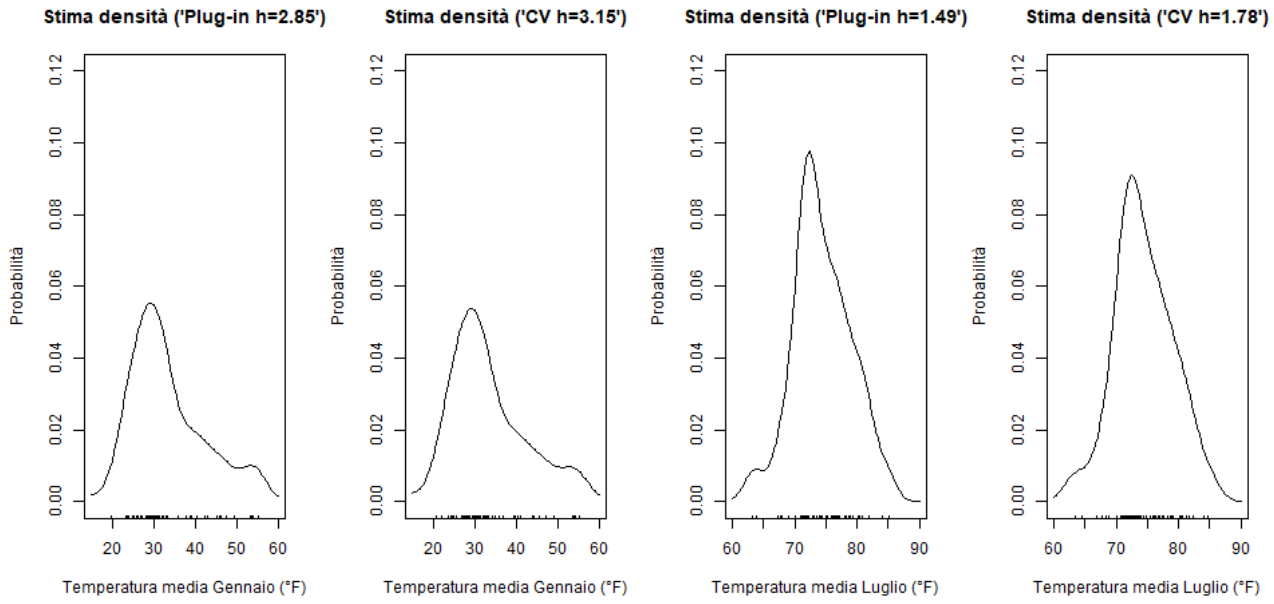
Inoltre, è possibile verificare la normalità anche mediante il qqplot che mette a confronto i quantili della distribuzione empirica con i quantili di una distribuzione normale standardizzata:




Dai grafici risulta più evidente la normalità nel primo caso e la non normalità nel secondo (i punti non seguono la retta).

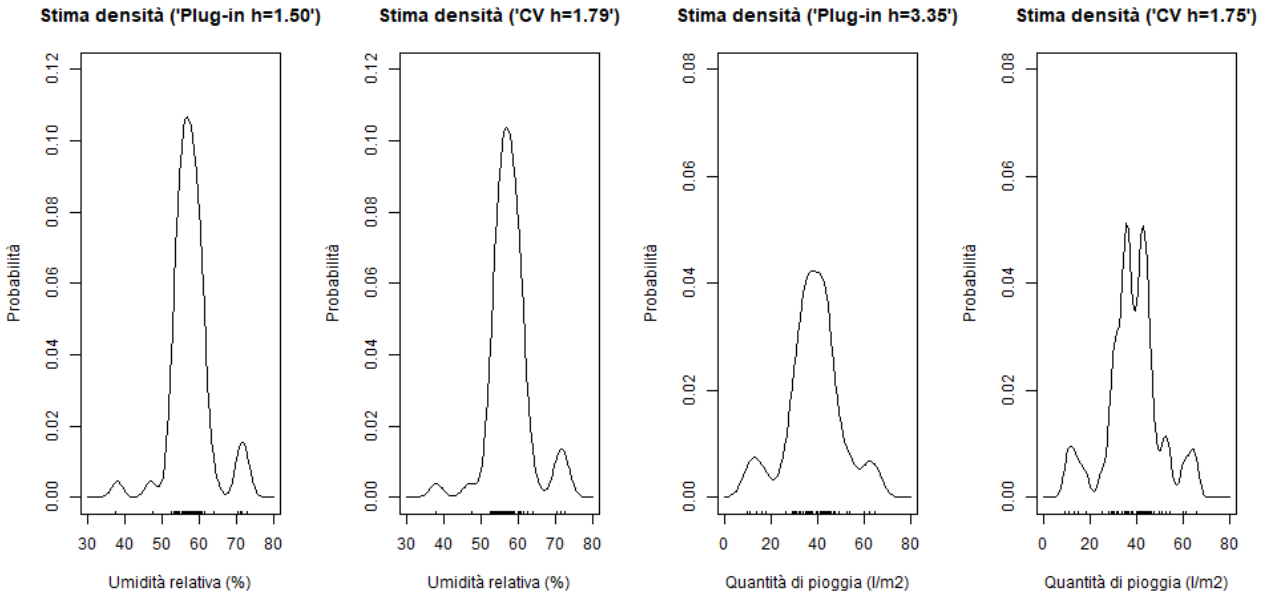
# Stimatori di nucleo univariati

Con gli istogrammi si era iniziato a vedere la forma della distribuzione, ossia della funzione di densità delle variabili (essendo variabili continue), però è conveniente stimare queste funzioni di densità attraverso gli 'stimatori di nucleo'; la stima di nucleo è regolata da un parametro $h$ chiamano 'parametro di smorzamento' che stabilisce la rugosità della funzione. Tale parametro può essere scelto arbitrariamente oppure scelto in modo automatico in base a 2 metodi: *Plug-in* e *Cross-Validation*.




In entrambi i casi il metodo della *Cross-Validation* assegna un valore al parametro $h$ più elevato ed infatti la distribuzione risulta essere più liscia rispetto al metodo *Plug-in*, tutto sommato però le stime risultano essere molto simili, unimodali in tutti i casi. La distribuzione relativa alle

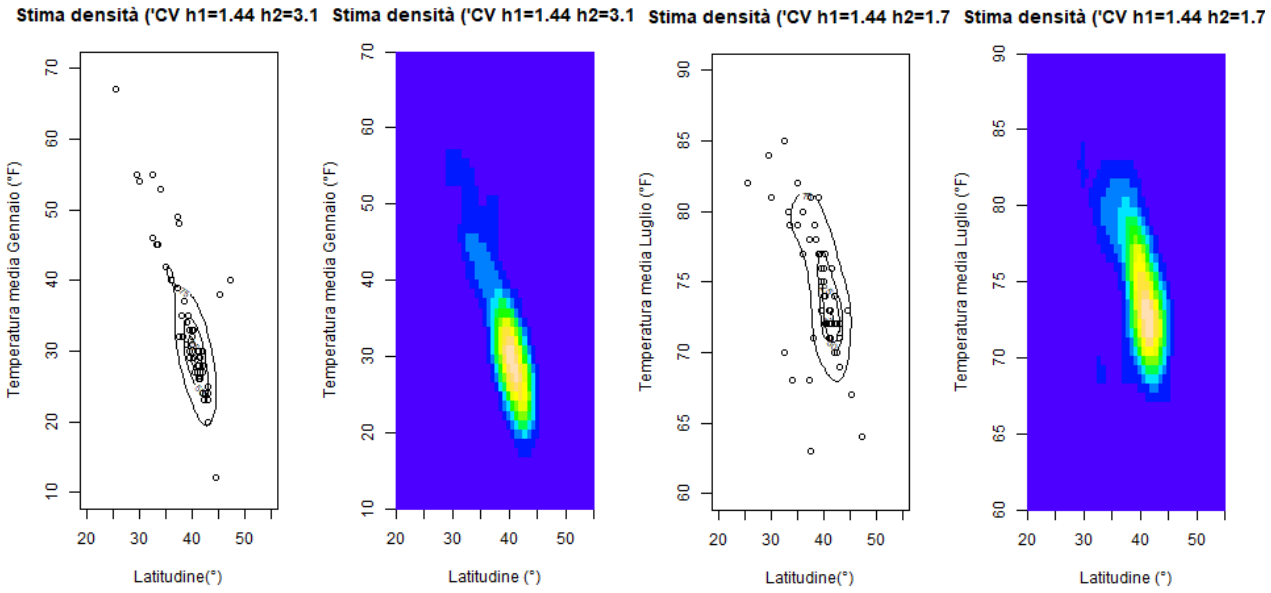
temperature del periodo estivo è più concentrata, cosa che già sapevamo dalla stima della varianza fatta in precedenza (101.31 per gennaio e 20.81 per luglio).



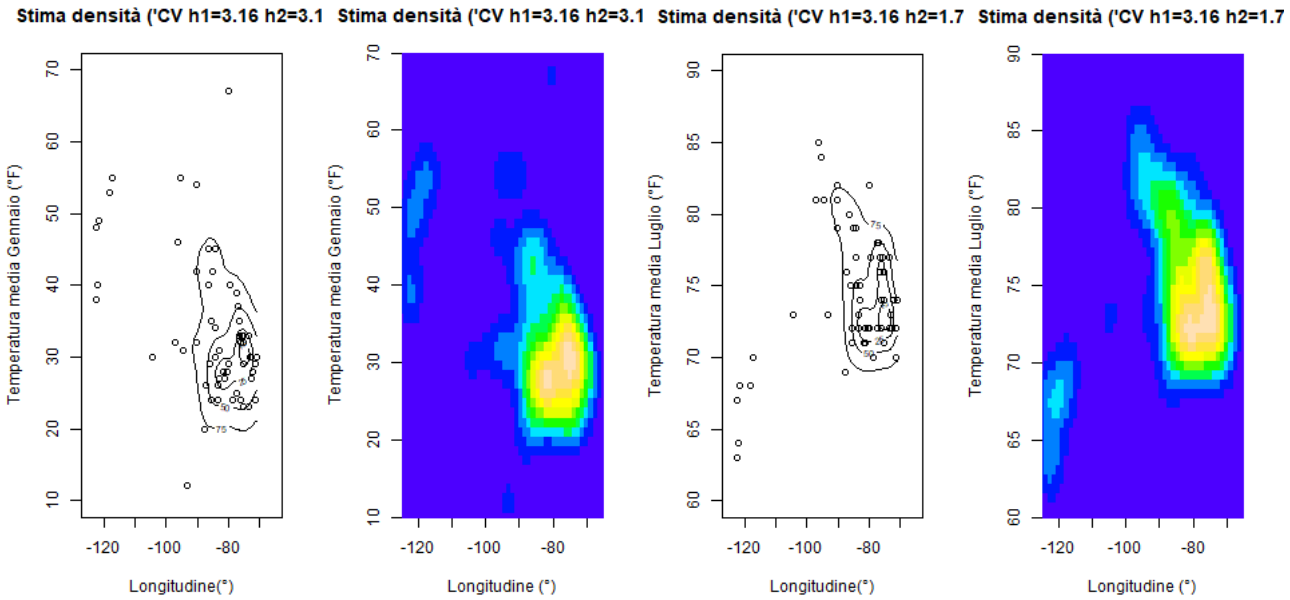

Per le altre 2 variabili quantitative oltre al picco modale centrale sembrano esserci altri picchi modali, seppur di minore intensità; anche qui, come sopra, il metodo CV assegna un parametro h più elevato con una conseguente stima della densità più liscia tranne che nell'ultimo caso dove, invece, il parametro assegnato con il metodo Plug-in è molto più elevato e in effetti la stima risulta essere molto più liscia, mentre con la CV risulta molto rugosa e tendente ad una distribuzione multimodale.

# Stimatori di nucleo bivariati

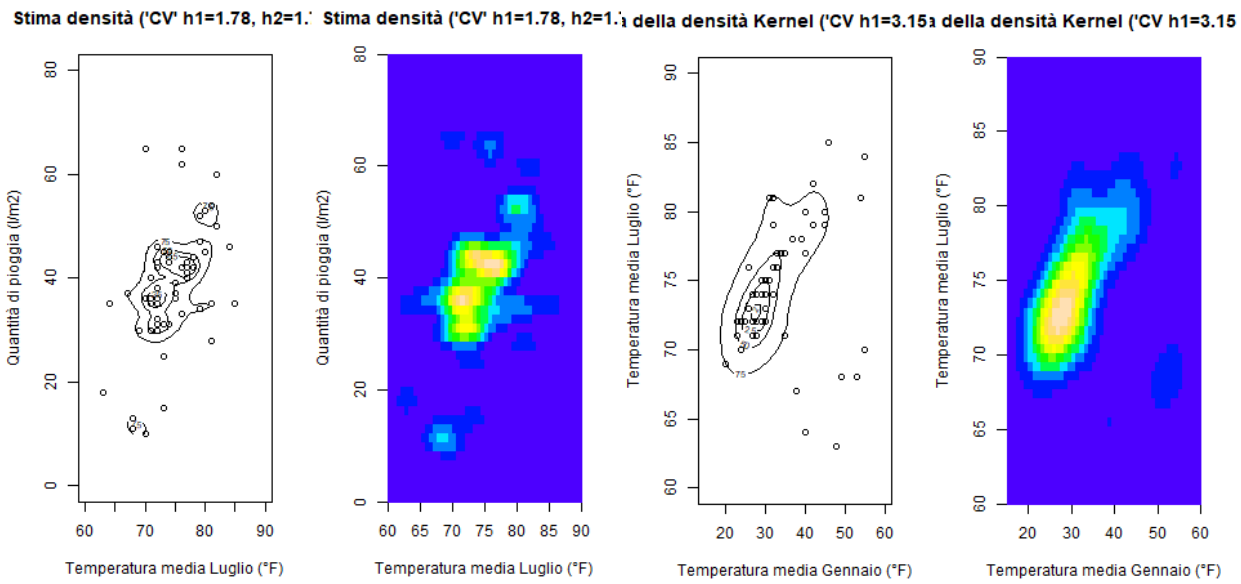
La stima di nucleo può essere fatta anche per coppie di variabili, attraverso modelli 3D o modelli che sfruttano le curve di livello. In questo caso sono stati presi in considerazione i secondi e per quanto riguarda la scelta automatica del parametro di smorzamento $h$ è stato usato il metodo della *CV*. I risultati ottenuti sono stati:




anche qui è possibile notare la dipendenza tra ‘latitudine’ e le temperature attraverso le ellissi che si formano (più schiacciato è più dipendenza c’è). In più, le curve di livello ci permettono di mostrare il 3D in 2D, assegnando delle curve a dei livelli di altezza; nel primo grafico notiamo che il picco massimo lo si ha intorno ai 40° di latitudine e 30 °F, mentre nel secondo grafico introno ai 42° di latitudine e 72°F.




Dai grafici notiamo che una dipendenza lineare c'è ma è poco intensa; la cosa più interessante è la divisione in 2 cluster, uno molto più accentuato dell'altro a causa della maggiore numerosità. Il cluster meno accentuato fa riferimento alla costa occidentale (-120° circa), dove abbiamo notato più volte esserci una relazione diversa rispetto al restante paese, e fa vedere molto bene il comportamento di quelle città più a Ovest di tutto, che in inverno tendono ad avere temperature molto elevate (e il restante paese temperature più basse) mentre in estate non raggiungono alte temperature.



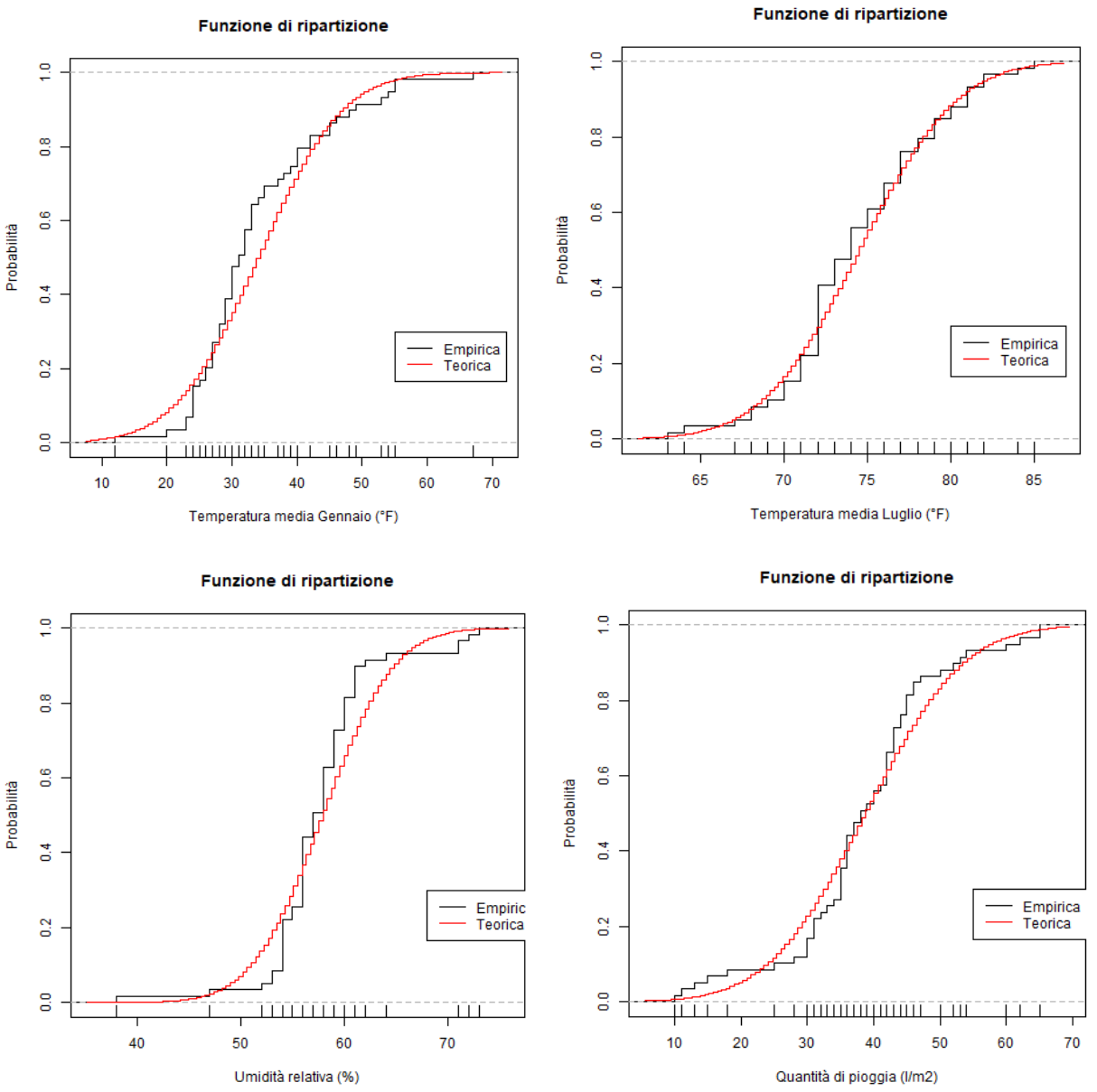

In questo penultimo grafico sembrano esserci più cluster (ovviamente la numerosità campionaria non molto elevata non aiuta) e tutto sommato una leggera dipendenza lineare positiva c'è; nell'ultimo grafico vi è un unico grande cluster che fa tendere ad una dipendenza positiva, ma considerando tutto quello che è stato visto finora è possibile che vi sia un'ulteriore cluster in corrispondenza di alti valori di JanTF e bassi valori di JulyTF (sempre per lo stesso ragionamento della costa occidentale che si comporta decisamente in modo diverso).

# Inferenza con una sola variabile

## Test di Kolmogorov

Dopo aver fatto un'attenta analisi esplorativa univariata e successivamente multivariata, andando a vedere bene quali relazioni ci fossero tra le variabili, si è passati ad un'analisi inferenziale, facendo delle ipotesi. La prima ipotesi che è stata fatta si riferiva non ad un determinato parametro

ma all'intera funzione di ripartizione. L'ipotesi di base $H_0$ prevede che la funzione di ripartizione empirica sia uguale ad una funzione di ripartizione teorica (nel nostro caso abbiamo scelto una distribuzione normale), quindi $H_0{:}F(x){=}F_0(x)$, contro l'ipotesi alternativa $H_1{:}F(x){\neq}F_0(x)$. Quest'ipotesi è chiamata 'Test di Kolmogorov' ed è stata fatta per tutte e 4 le variabili:




Dalla visualizzazione grafica sembra che tutte le variabili seguano una distribuzione normale, in particolar modo la variabile JulyTF e Rain.

```
> ks.test(JanTF, "pnorm", 33.8, 10.15)

        One-sample Kolmogorov-Smirnov test

data:  JanTF
D = 0.17548, p-value = 0.05284
alternative hypothesis: two-sided

> ks.test(RelHum, "pnorm", 57.75, 5.37)

        One-sample Kolmogorov-Smirnov test

data:  RelHum
D = 0.17082, p-value = 0.06391
alternative hypothesis: two-sided
```

```
> ks.test(JulyTF, "pnorm", 74.4, 4.6)

        One-sample Kolmogorov-Smirnov test

data:  JulyTF
D = 0.10585, p-value = 0.523
alternative hypothesis: two-sided

> ks.test(Rain, "pnorm", 38.5, 11.57)

        One-sample Kolmogorov-Smirnov test

data:  Rain
D = 0.11263, p-value = 0.4426
alternative hypothesis: two-sided
```

L'implementazione del test conferma quanto detto, dando un *p-value* in ogni caso maggiore di 0.05 (livello di significatività 5%), soprattutto per quanto riguarda la variabile JulyTF e Rain.

## Test di Mann-Whitney

Successivamente è stata creata una variabile *dummy* in base alla longitudine (0 se ≤ -83.00°, 1 se > -83.00°). Ciò è stato fatto per poter vedere se le mediane delle temperature di gennaio e luglio fossero differenti significativamente nei 2 campioni attraverso un 'Test di Mann-Whitney' che suppone $H_0{:}\lambda_1 = \lambda_2$, contro $H_1{:}\lambda_1 \neq \lambda_2$ (ipotesi bilaterale) oppure $H_0{:}\lambda_1 = \lambda_2$, contro $H_1{:}\lambda_1 > \lambda_2$ (ipotesi direzionale). Il test è 'distribution-free' ed è stato preferito al 'Test T-student' per le forti assunzioni che quest'ultimo comporta come la normalità dei due campioni e l'omoschedasticità.

È stato implementato il test bilaterale per la variabile JulyTF e il test direzionale per la variabile JanTF:

```
> wilcox.test(JanTF ~ Dummy_Long, alternative = "greater")

        Wilcoxon rank sum test with continuity correction

data:  JanTF by Dummy_Long
W = 584.5, p-value = 0.01184
alternative hypothesis: true location shift is greater than 0

> wilcox.test(JulyTF ~ Dummy_Long, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  JulyTF by Dummy_Long
W = 496, p-value = 0.3568
alternative hypothesis: true location shift is not equal to 0
```

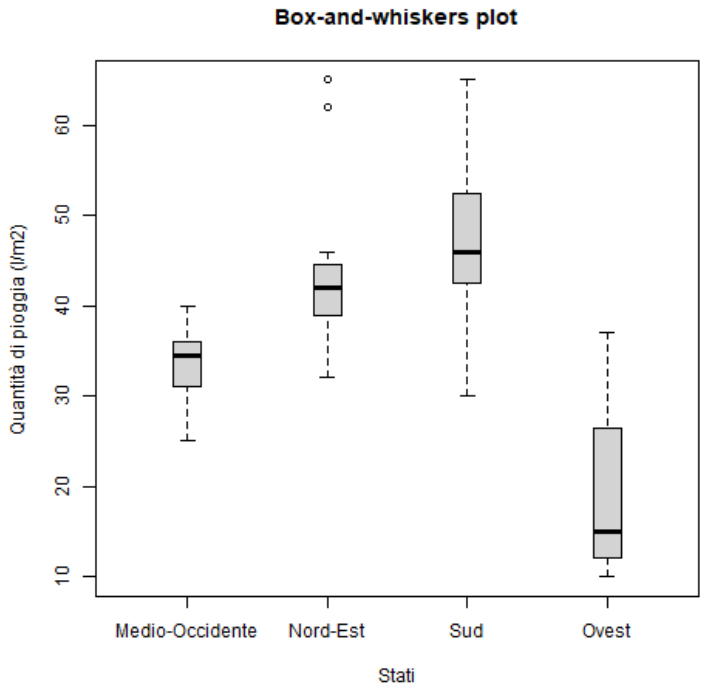
Il primo test ci dà un *p-value* pari a 0.012, molto basso, che ci fa rifiutare l'ipotesi di base a favore dell'ipotesi alternativa $H_1{:}\lambda_1 > \lambda_2$, quindi il valore mediano delle temperature di gennaio delle città più a Ovest è maggiore del valore mediano delle temperature di gennaio delle città più a Est (> -83° longitudine).

Il secondo test invece, con un *p-value* pari a 0.36 ci fa accettare l'ipotesi di base, ossia l'uguaglianza delle mediane, per quanto riguarda le temperature di luglio.

## Test di Kruskal-Wallis

Riprendendo la variabile qualitativa a 4 livelli State, è stato fatto anche qui un test per verificare l'ipotesi di uguaglianza delle mediane in ciascun gruppo; essendo i gruppi maggiori di 2 non si può implementare un 'Test di Mann-Whitney' bensì un 'Test di Kruskal-Wallis'.

Non è stata implementata un'analisi della varianza per verificare l'uguaglianza delle medie per le sue assunzioni forti (normalità nei gruppi). La variabile presa in considerazione questa volta è stata Rain, analizzando prima il boxplot.




Il boxplot mostra delle differenze tra le mediane, in particolar modo quella relativa alle città occidentali è molto più bassa rispetto alle altre, ciò che si vuole vedere è se tale differenza è statisticamente significativa oppure no:

```
> kruskal.test(Rain ~ State)

        Kruskal-Wallis rank sum test

data:  Rain by State
Kruskal-Wallis chi-squared = 33.341, df = 3, p-value = 2.729e-07
```

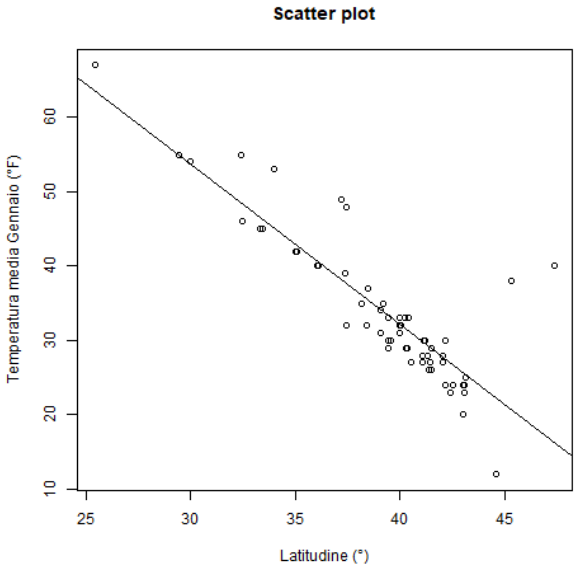
Il test ci fornisce un *p-value* che tende molto a 0, quindi ci fa rifiutare fortemente l'ipotesi di base di uguaglianza delle mediane, il che vuol dire che almeno una mediana differisce dalle altre (probabilmente sarà quella relativa ad Ovest).

# Regressione

Una volta visto che tra alcune variabili evidentemente c'è una dipendenza lineare si è passati all'implementazione di un 'modello di regressione' per cercare di prevedere una variabile in funzione di un'altra. Si è considerato il modello di regressione $Y_i = \beta_0+\beta_1 x_1+\varepsilon_i$ dove gli errori $\varepsilon_i$ hanno media 0 e varianza $\sigma^2$. I parametri $\beta_0$ e $\beta_1$ vengono stimati attraverso il metodo OLS. Considerando un modello con variabile dipendente JanTF e variabile indipendente la Latitudine abbiamo la stima dei parametri:

```
> lm(JanTF ~ Lat)

Call:
lm(formula = JanTF ~ Lat)

Coefficients:
(Intercept)          Lat
     118.14        -2.15
```

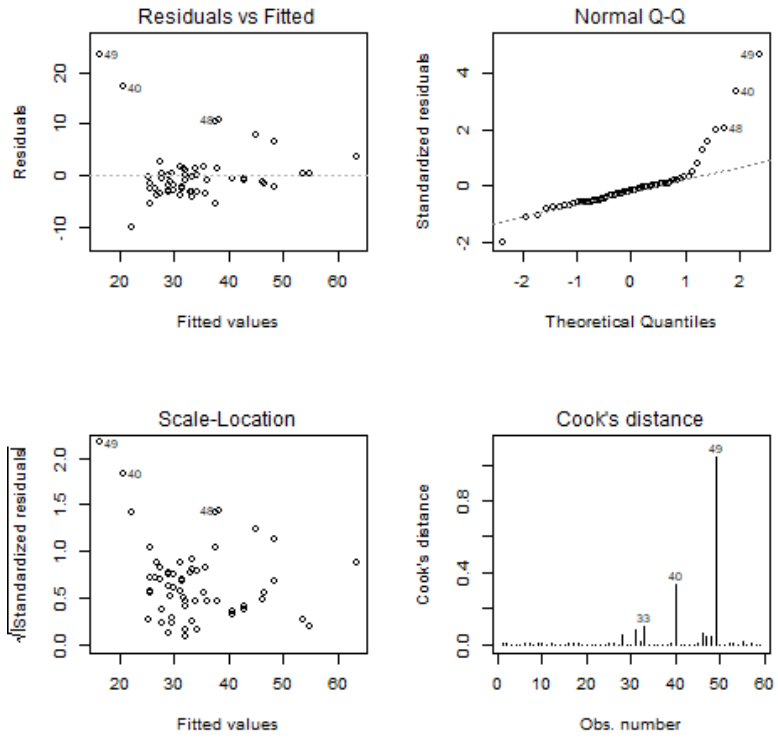
Il relativo grafico con la retta di regressione stimata è invece:



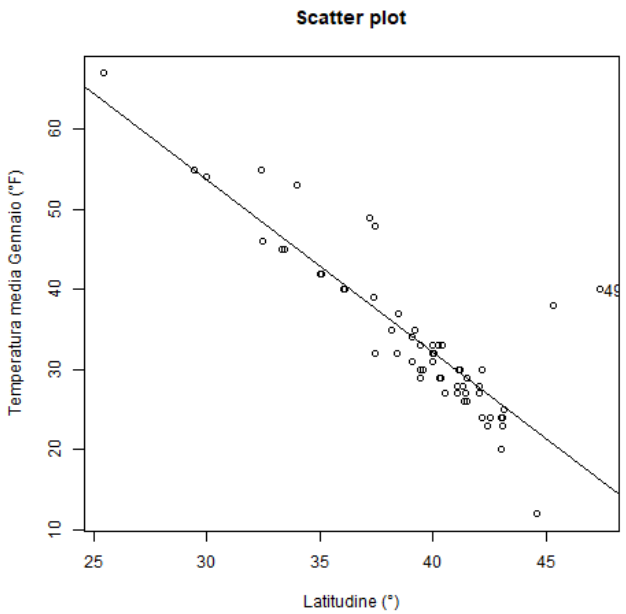

```
> summary(Reg)

Call:
lm(formula = JanTF ~ Lat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2978  -2.6353  -0.8719   0.3965  23.6789

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  118.139      6.743   17.52   <2e-16 ***
Lat           -2.150      0.171  -12.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.272 on 57 degrees of freedom
Multiple R-squared:  0.735,     Adjusted R-squared:  0.7303
F-statistic: 158.1 on 1 and 57 DF,  p-value: < 2.2e-16
```

Andando ad analizzare il modello abbiamo che, sia l'intercetta e il coefficiente angolare risultato statisticamente significativi, inoltre il modello è piuttosto buono poiché $R^2$=0.735.

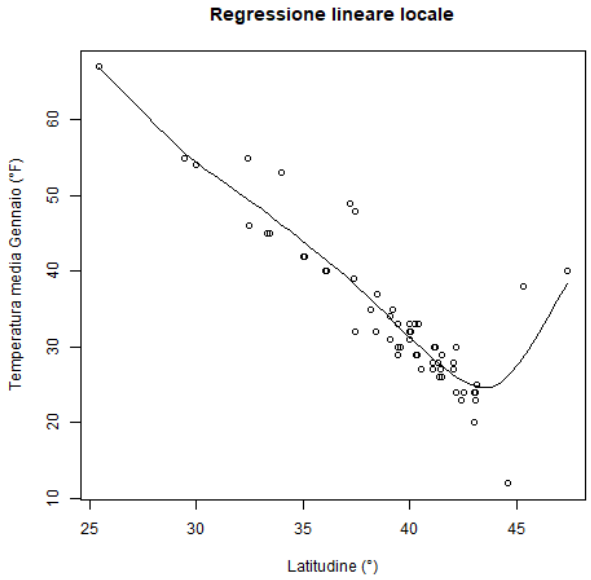
Oltre al modello è importante vedere e analizzare i residui, ciò lo si fa attraverso 4 grafici:




Da questi grafici possiamo vedere l'influenza delle singole osservazioni e vediamo in particolar modo che l'unità statistica 49 (Seattle) influisce negativamente e, infatti, riprendendo lo scatter plot vediamo:



Andando a fare una ‘regressione lineare locale’ dove $h$ è scelto attraverso il metodo dei *gradi di libertà* abbiamo:




Dalla stima di questo modello di regressione lineare locale è ancora più evidente la distorsione che da l’unità statistica 49 quindi abbiamo provato ad implementare il modello senza considerare tale osservazione.

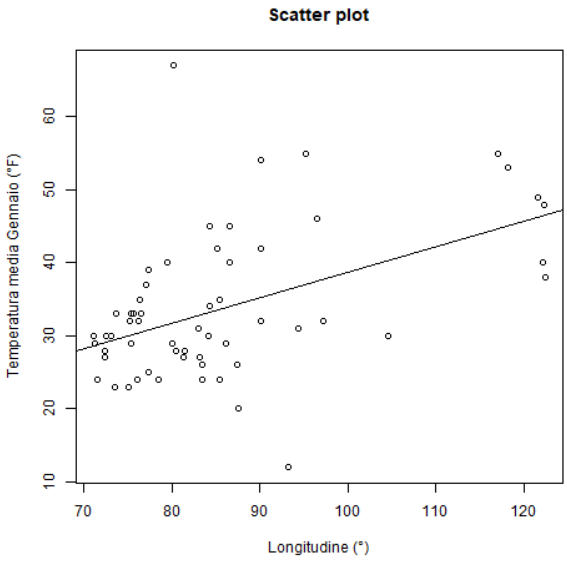
```
> summary(lm(JanTF ~ Lat, subset = (1:length(JanTF) !=  49)))

Call:
lm(formula = JanTF ~ Lat, subset = (1:length(JanTF) != 49))

Residuals:
    Min      1Q  Median      3Q     Max
-8.6729 -2.0620 -0.8525  1.0969 19.0583

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.3950     5.5023   22.97   <2e-16 ***
Lat          -2.3715     0.1401  -16.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.163 on 56 degrees of freedom
Multiple R-squared:  0.8366,    Adjusted R-squared:  0.8337
F-statistic: 286.7 on 1 and 56 DF,  p-value: < 2.2e-16
```

Il risultato è stato che togliendo l’osservazione il modello migliora, passando da un $R^2$=0.735 ad un $R^2$=0.837.

Abbiamo anche provato ad implementare un modello di regressione usando come variabile indipendente la Longitudine:

```
> lm(JanTF ~ Long)

Call:
lm(formula = JanTF ~ Long)

Coefficients:
(Intercept)         Long
     4.0299       0.3473
```

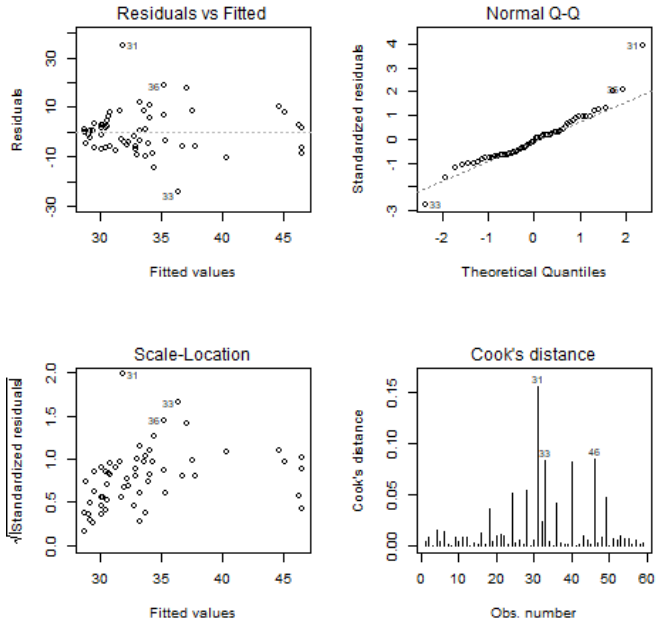
Il grafico con la retta stimata è:



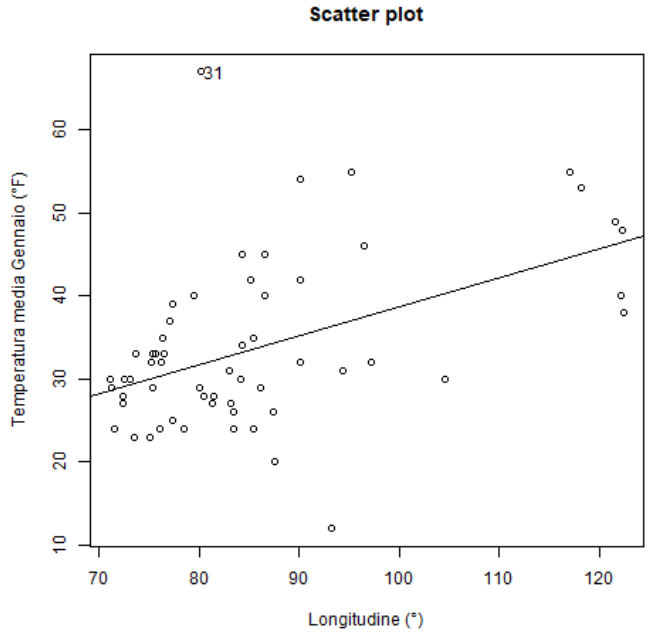

```
> summary(Reg1)

Call:
lm(formula = JanTF ~ Long)

Residuals:
    Min      1Q  Median      3Q     Max
-24.378  -5.838  -1.155   3.932  35.151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.02990    7.38093   0.546 0.587204
Long         0.34727    0.08501   4.085 0.000139 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.007 on 57 degrees of freedom
Multiple R-squared:  0.2264,    Adjusted R-squared:  0.2129
F-statistic: 16.69 on 1 and 57 DF,  p-value: 0.0001394
```

Come prima cosa vediamo che l’ $R^2$ è molto basso, 0.22, il modello quindi non è molto buono, inoltre risulta statisticamente significativo solo il coefficiente angolare e non l’intercetta.

Per vedere come le unità statistiche contribuiscono al modello implementiamo un'analisi dei residui:

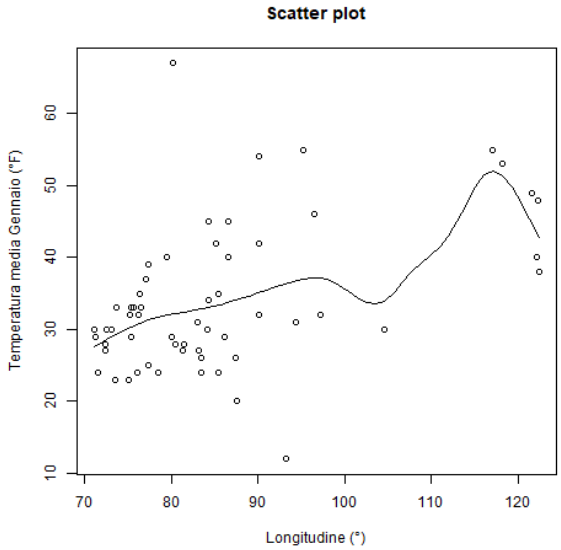



Dal diagramma risulta che l'unità 31 (Miami) potrebbe influenzare negativamente il modello, quindi conviene vedere qual è sullo scatter plot.




In effetti l'unità 31 sembrerebbe essere molto anomala, un residuo molto elevato che quindi potrebbe influenzare il modello.

Passiamo così ad una ‘regressione lineare locale’:




Dal grafico risulta una relazione abbastanza lineare nella prima parte (con residui abbastanza elevati), mentre nella seconda parte non vi è una buona relazione, magari dovuto anche alle poche osservazioni riguardo la costa occidentale.

Avendo visto che l’osservazione 31 può causare problemi al modello si è deciso di provare a fare la regressione escludendo tale osservazione.

```
> summary(lm(JanTF ~ Long, subset = (1:length(JanTF) != 31)))

Call:
lm(formula = JanTF ~ Long, subset = (1:length(JanTF) != 31))

Residuals:
     Min       1Q   Median       3Q      Max
-23.9032  -5.3125  -0.3461   4.0836  19.2289

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.88664    6.36824   0.296    0.768
Long         0.36518    0.07326   4.985 6.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.75 on 56 degrees of freedom
Multiple R-squared:  0.3074,    Adjusted R-squared:  0.295
F-statistic: 24.85 on 1 and 56 DF,  p-value: 6.321e-06
```

Togliendo l'osservazione 31 in effetti il modello è migliorato da $R^2$=0.22 a $R^2$=0.30, però l'intercetta risulta ancora non significativa, quindi è stato provato un modello senza intercetta.

```
> summary(lm(JanTF ~ -1 + Long, subset = (1:length(JanTF) != 31)))

Call:
lm(formula = JanTF ~ -1 + Long, subset = (1:length(JanTF) !=
    31))

Residuals:
    Min      1Q  Median      3Q     Max
-24.012  -5.341  -0.041   4.385  19.186

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
Long  0.38660    0.01161   33.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.688 on 57 degrees of freedom
Multiple R-squared:  0.9511,    Adjusted R-squared:  0.9502
F-statistic:  1108 on 1 and 57 DF,  p-value: < 2.2e-16
```

Togliendo l'intercetta il modello migliora davvero molto arrivando ad un $R^2$=0.95, valore molto elevato.

# Regressione multipla

Per ultimo è stato implementato un modello che contenesse come regressori tutte le variabili e come variabile di risposta la variabile JulyTF:

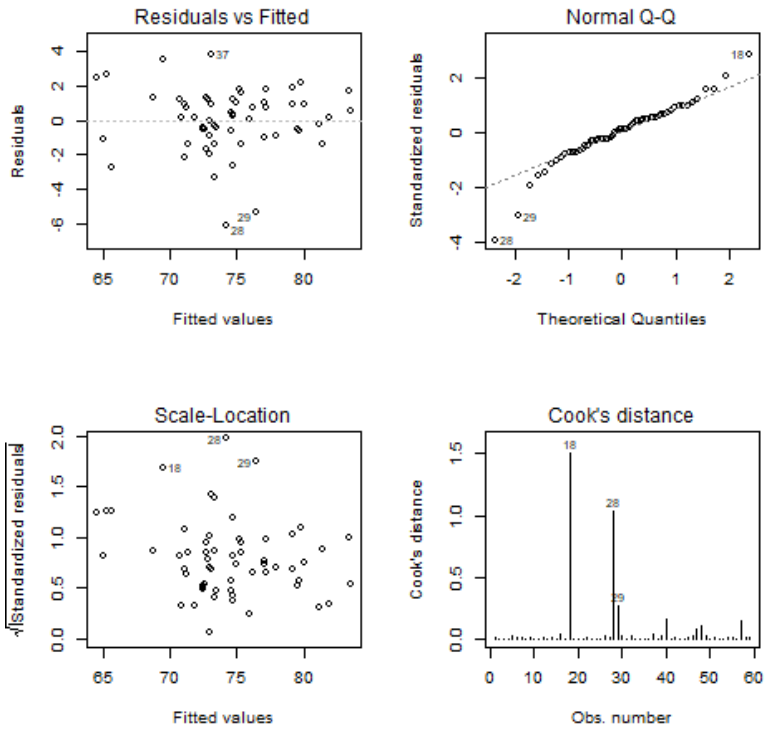
```
> summary(lm(JulyTF ~ Lat + Long + JanTF + Rain + RelHum + State))

Call:
lm(formula = JulyTF ~ Lat + Long + JanTF + Rain + RelHum + State)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2331 -0.9050  0.2472  1.0739  3.8194

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     60.65654   11.92433   5.087 5.50e-06 ***
Lat              0.05857    0.24267   0.241   0.8103
Long             0.30678    0.06258   4.902 1.04e-05 ***
JanTF            0.21173    0.10878   1.946   0.0572 .
Rain             0.07117    0.03876   1.836   0.0723 .
RelHum          -0.39608    0.07166  -5.527 1.18e-06 ***
StateNorth-East  0.58133    0.93460   0.622   0.5368
StateSouth       0.33802    1.07416   0.315   0.7543
StateWest      -18.05003    2.65747  -6.792 1.27e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.968 on 50 degrees of freedom
Multiple R-squared:  0.8423,    Adjusted R-squared:  0.8171
F-statistic: 33.39 on 8 and 50 DF,  p-value: < 2.2e-16
```

Coinvolgendo tutte le variabili nel modello risulta un $R^2$=0.81, ma non tutte le variabili risultano significative, per questo si è voluto adottare il 'criterio di Akaike' che rende il modello più semplice andando a togliere alcuni regressori.

```
> summary(step(lm(JulyTF ~ Lat + Long + JanTF + Rain + RelHum + State)))
Start:  AIC=88.13
JulyTF ~ Lat + Long + JanTF + Rain + RelHum + State

         Df Sum of Sq    RSS     AIC
- Lat     1     0.226 193.89  86.195
<none>                193.66  88.127
- Rain    1    13.057 206.72  89.976
- JanTF   1    14.674 208.34  90.436
- Long    1    93.083 286.75 109.283
- RelHum  1   118.327 311.99 114.261
- State   3   205.153 398.82 124.747

Step:  AIC=86.2
JulyTF ~ Long + JanTF + Rain + RelHum + State

         Df Sum of Sq    RSS     AIC
<none>                193.89  86.195
- Rain    1     15.56 209.45  88.751
- JanTF   1     72.00 265.89 102.827
- Long    1     95.18 289.07 107.760
- RelHum  1    174.37 368.26 122.044
- State   3    431.62 625.51 149.300

Call:
lm(formula = JulyTF ~ Long + JanTF + Rain + RelHum + State)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1501 -0.9152  0.1934  1.1048  3.8587

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     63.26575    4.98429  12.693  < 2e-16 ***
Long             0.30370    0.06070   5.004 7.06e-06 ***
JanTF            0.18767    0.04313   4.352 6.50e-05 ***
Rain             0.07402    0.03658   2.023   0.0483 *
RelHum          -0.38576    0.05696  -6.772 1.25e-08 ***
StateNorth-East  0.60139    0.92226   0.652   0.5173
StateSouth       0.35192    1.06266   0.331   0.7419
StateWest      -17.62812    1.98298  -8.890 6.08e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.95 on 51 degrees of freedom
Multiple R-squared:  0.8421,    Adjusted R-squared:  0.8205
F-statistic: 38.87 on 7 and 51 DF,  p-value: < 2.2e-16
```

Il criterio di Akaike suggerisce di togliere la variabile Latitudine facendo passare l'$R^2$ da 0.817 a 0.82, un piccolo miglioramento, inoltre anche il valore AIC migliora passando da 88.13 a 86.2.

Volendo vedere l'analisi dei residui implementiamo i 4 grafici:




Dal diagramma sembra che l'unità 28 (Los Angeles) sia un valore anomalo per il modello, quindi si è provato ad implementare il modello togliendo tale unità.

```
> summary(step(lm(JulyTF ~ Long + JanTF + Rain + RelHum + State, subset = (1:length(JulyTF) !=  28))))
Start:  AIC=65.26
JulyTF ~ Long + JanTF + Rain + RelHum + State

         Df Sum of Sq    RSS     AIC
<none>                135.59  65.255
- Rain    1     10.40 146.00  67.543
- JanTF   1    100.83 236.43  95.502
- Long    1    107.30 242.89  97.067
- RelHum  1    230.63 366.23 120.883
- State   3    386.64 522.23 137.465

Call:
lm(formula = JulyTF ~ Long + JanTF + Rain + RelHum + State, subset = (1:length(JulyTF) !=
    28))

Residuals:
    Min      1Q  Median      3Q     Max
-5.1769 -0.8630  0.1040  0.9783  3.8068

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     66.90219    4.28210  15.624  < 2e-16 ***
Long             0.32358    0.05144   6.290 7.73e-08 ***
JanTF            0.22849    0.03747   6.098 1.54e-07 ***
Rain             0.06078    0.03103   1.959   0.0557 .
RelHum          -0.48824    0.05294  -9.222 2.28e-12 ***
StateNorth-East  0.70710    0.77926   0.907   0.3685
StateSouth      -0.36600    0.91077  -0.402   0.6895
StateWest      -17.48548    1.67508 -10.439 3.72e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.647 on 50 degrees of freedom
Multiple R-squared:  0.8857,    Adjusted R-squared:  0.8697
F-statistic: 55.36 on 7 and 50 DF,  p-value: < 2.2e-16
```

Togliendo l'unità 28 effettivamente il modello migliora passando da $R^2$ pari 0.82 a 0.87 ed anche l'AIC migliora passando da 86.2 a 65.26.

In conclusione possiamo dire che il comportamento delle temperature negli USA a gennaio e a luglio non è lo stesso per tutto il paese; nella parte ad ovest del paese le temperature restano abbastanza costanti durante tutto l'anno, con inverni piuttosto caldi ed estati che non raggiungono temperature elevatissime (dal Canada in giù inoltre scendono ancora di più); nella parte a nord-est e nel medio-occidente le temperature sono molto omogenee sia in estate che in inverno ma con temperature decisamente più alte in estate (varianza within molto bassa e varianza between molto elevata); infine, nella parte a sud del paese si hanno temperature molto elevate sia in inverno che in estate dovute alla vicinanza del Tropico del Cancro. La variabile che riguarda la temperatura a gennaio, inoltre, può essere stimata molto bene con un modello di regressione lineare semplice che considera come regressore la latitudine dando un $R^2$=0.837; se si usa invece un modello di regressione che considera come regressore la longitudine allora si ha un modello ancora più efficace, senza intercetta, che spiega addirittura il 95% della variabilità.

Per ultimo, volendo stimare la temperatura a luglio e usando come regressori tutte le variabili a disposizione si è ottenuto un modello che raggiunge un $R^2$ molto buono pari a 0.87.