In [146]:
```
using DataFrames, RDatasets, Base, Distributions;
using Plots;
gr()
```

Out[146]:
```
Plots.GRBackend()
```

Question 1

a. Load dataset

In [161]:
```
data = dataset("car","Davis")
```

Out[161]:

	Sex	Weight	Height	RepWt	RepHt
1	M	77	182	77	180
2	F	58	161	51	159
3	F	53	161	54	158
4	M	68	177	70	175
5	F	59	157	59	155
6	M	76	170	76	165
7	M	76	167	77	165
8	M	69	186	73	180
9	M	71	178	71	175
10	M	65	171	64	170
11	M	70	175	75	174
12	F	166	57	56	163

In [97]:
```
1 height_data, weight_data = data[:, [:Height]], data[:, [:Weight]]
```

Out[97]:

(200×1 DataFrames.DataFrame

Row	Height
1	182
2	161
3	161
4	177
5	157
6	170
7	167
8	186
9	178
10	171
11	175
⋮	
189	183
190	158
191	185
192	173
193	164
194	156
195	164
196	175
197	180
198	175
199	181
200	177

, 200×1 DataFrames.DataFrame

Row	Weight
1	77
2	58
3	53
4	68
5	59
6	76
7	76
8	69
9	71
10	65
11	70
⋮	
189	76
190	50
191	88
192	89
193	59
194	51
195	62
196	74
197	83
198	81
199	90
200	79

)

b. Find mean of Weight and Height

Below are the equations for estimating the mean and variance of the two variables 'Height' and 'Weight', treating them as univariate variables

$$\hat{\mu}_{Height} = \frac{\sum_{i=1}^{N} x_i^{Height}}{N}$$

$$\hat{\mu}_{Weight} = \frac{\sum_{i=1}^{N} x_i^{Weight}}{N}$$

$$\hat{\sigma}_{Height} = \frac{\sum_{i=1}^{N} (x_i^{Height} - \hat{\mu}_{Height})^2}{N}$$

$$\hat{\sigma}_{Weight} = \frac{\sum_{i=1}^{N} (x_i^{Weight} - \hat{\mu}_{Weight})^2}{N}$$

In [127]:
```
function get_mean(data)
    nrow, ncol = size(data);
    return sum(data, 1)./nrow
end
```

Out[127]:
```
get_mean (generic function with 1 method)
```

In [137]:
```
function get_variance(data)
    nrow, _ = size(data);
    mean_vector = get_mean(data)
    return sum(( data - repmat(mean_vector, nrow) ).^2, 1)./nrow
end
```

Out[137]:
```
get_variance (generic function with 1 method)
```

In [134]:
```
data_1b = data[:, [:Height, :Weight] ];
data_1b = convert(Array, data_1b);
```

In [136]:
```
height_mean, weight_mean = get_mean(data_1b);
print("Height_mean: ", height_mean, "\n");
print("Weight_mean: ", weight_mean, "\n");
```

```
Height_mean: 170.02
Weight_mean: 65.8
```

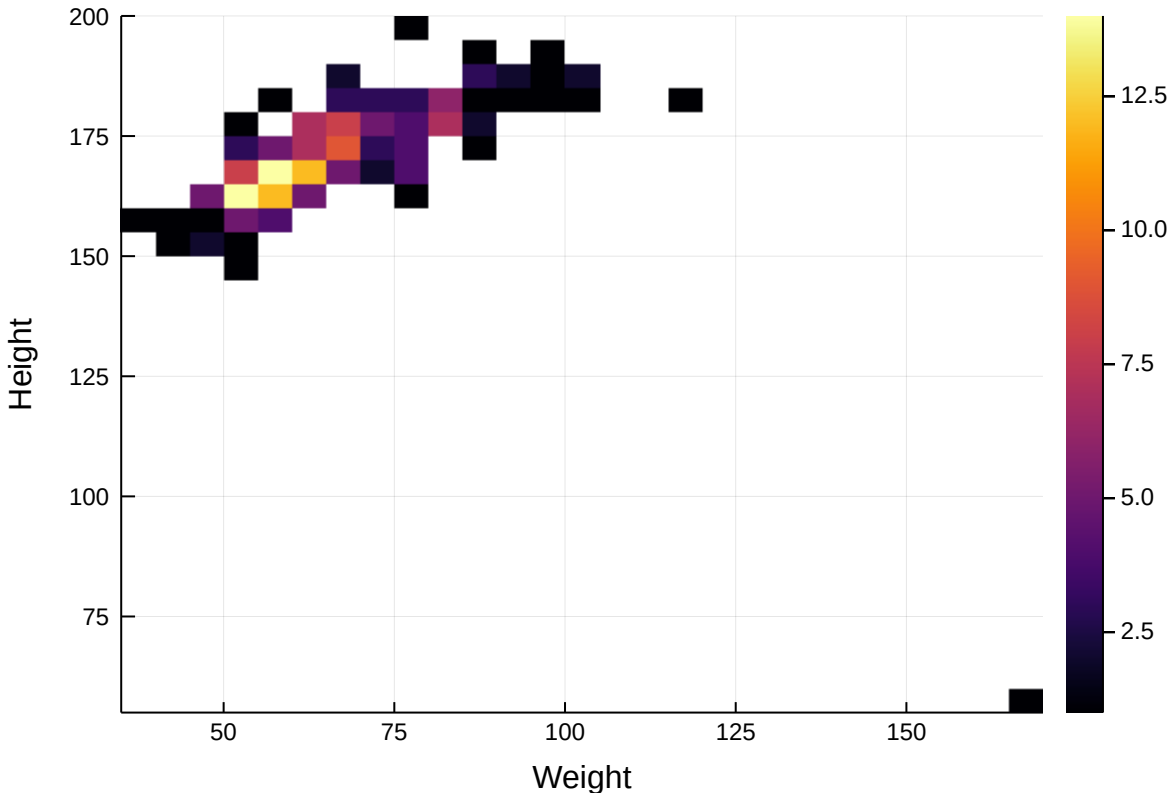
In [138]:
```
height_variance, weight_variance = calc_variance(data_1b)
print("Height_variance: ", height_variance, "\n");
print("Weight_variance: ", weight_variance, "\n");
```

```
Height_variance: 143.46959999999996
Weight_variance: 226.71999999999997
```

c. Draw histogram

In [43]:
```
histogram2d(data[:Weight],data[:Height],nbins=40,xlabel="Weight",ylab
```

Out[43]:



Observations from the above plot:

1. Weight and Height data are dependent.
2. Data shows a positive covariance.
3. It can be seen that as weight increases there is a increase in height.
4. There are few outliers in the dataset with weight around 200 and height around 50.

d. Mean and Covariance Estimate for Multivariate Case

Equations for estimating the mean vector and covariance matrix assuming they follow a multivariate Gaussian distribution.

x_i is a vector as follows: $\begin{bmatrix} x_i^{Height} & x_i^{Weight} \end{bmatrix}^T$

$\hat{\mu}$ is a vector as follows: $\begin{bmatrix} \hat{\mu}_{Height} & \hat{\mu}_{Weight} \end{bmatrix}^T$

Equation for MEAN VECTOR ESTIMATE:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Equation for COVARIANCE MATRIX ESTIMATE:

$$\hat{\Sigma} = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu})(x_i - \hat{\mu})^T}{N} = \begin{bmatrix} \sigma_H^2 & \sigma_{HW}^2 \\ \sigma_{HW}^2 & \sigma_W^2 \end{bmatrix}$$

In [140]:
```
function covariance_multiVariate(data)
    nrow, _ = size(data);
    mean_vector = get_mean(data);
    return ((data_matrix - repmat(mean_vector, nrow))'*(data_matrix -
end
```

Out[140]:
```
covariance_multiVariate (generic function with 2 methods)
```

In [162]:
```
mean_vector = get_mean(data_1b);
print("Estimate of Mean Vector: [mean_height, mean_weight] = ", mean_
```

```
Estimate of Mean Vector: [mean_height, mean_weight] = [170.02 65.8]
```

In [163]:
```
cov_mat = covariance_multiVariate(data_1b);
print("Estimate of Covariance Matrix: \n")
print(cov_mat);
```

```
Estimate of Covariance Matrix:
[143.47 34.204; 34.204 226.72]
```

e. Julia code for distribution using estimated parameters

In [154]:
```
multi_gaussian_distribution = MvNormal(vec(mean_vector), cov_mat);
```

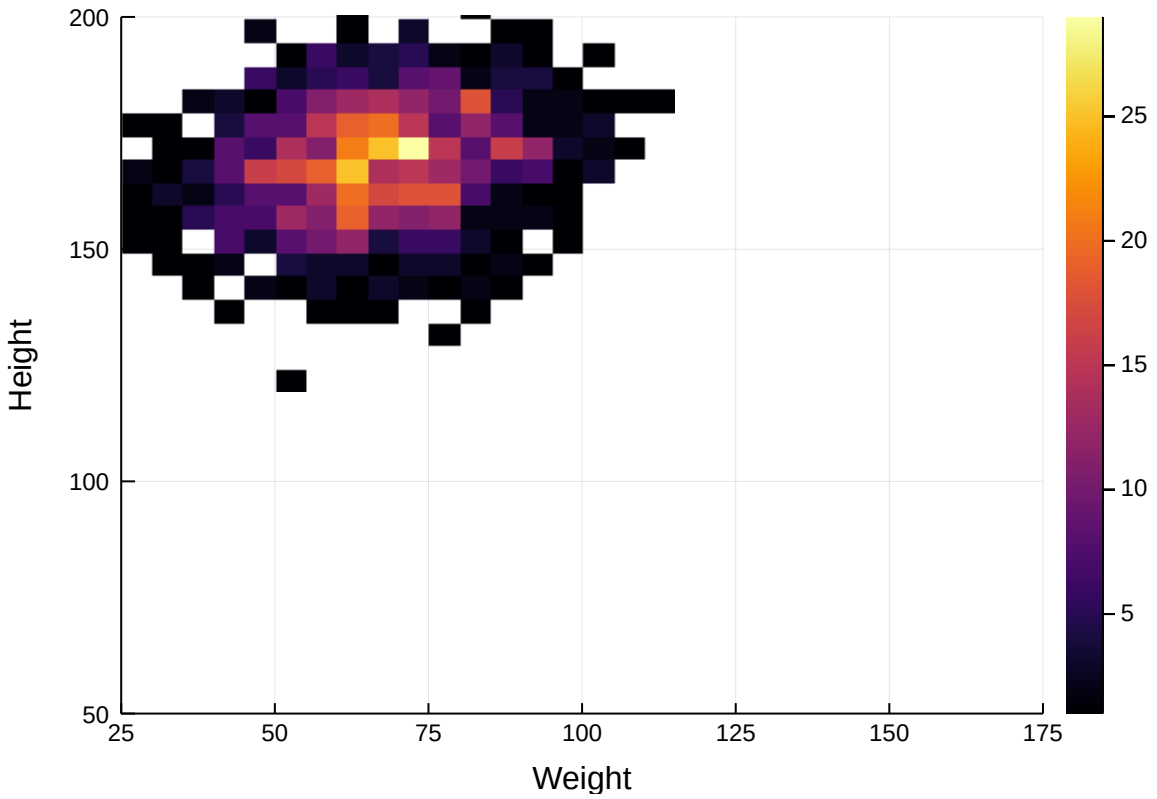
Out[154]:
```
FullNormal(
dim: 2
μ: [170.02, 65.8]
Σ: [143.47 34.204; 34.204 226.72]
)
```

In [155]:
```
1 data_samples = rand(multi_gaussian_distribution, 1000)
```

Out[155]:
```
2×1000 Array{Float64,2}:
 172.721   170.737   171.557  148.873   …  179.8     178.71    173.509
  85.0369   73.8594   51.44    65.7229      61.0291   84.7707   64.7466
```

In [160]:
```
1 histogram2d(data_samples[2, :], data_samples[1, :], nbins=40, xlabel=
```

Out[160]:



Comparing plots in question (1.c) and (1.e)

1. In the plot from (1.e), it can be seen that the covariance is small as compared to the plot from (1.c)
2. The data appears to be linearly dependent in (1.c) while in the plot generated by (1.e), it can be seen that the interdependency of the data is very less, i.e., covariance values are small.
3. There are few outlier in the data (1.c).
4. Estimated multivariate distribution, (1.e), does not appear to be a good fit to the data as the covariance observed is very less as compared to the given data.

f. From the covariance matrix, determine if the variables 'Height' and "Weight" are independent.

It can be seen from the covariance matrix that the off diagonal values are non-zero. Therefore, the data is dependent.

In [ ]:
```
1
```