Data Cleaning

In [66]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.show()
%matplotlib inline
```

In [53]:
```
databytime = pd.read_csv("databytime.csv")
databymonth = pd.read_csv("databymonth.csv")
```

In [518]:
```
databymonth.head()
```

Out[518]:

	STATE/UT	YEAR	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGU
0	A & N Islands	2001	8	23	15	15	14	19	14	19
1	A & N Islands	2002	12	10	14	16	10	7	16	11
2	A & N Islands	2003	19	13	15	13	13	12	8	16
3	A & N Islands	2004	21	14	22	17	13	18	16	19
4	A & N Islands	2005	19	21	22	17	13	19	21	14

In [41]:
```
databymonth.tail()
```

Out[41]:

	STATE/UT	YEAR	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AU
485	West Bengal	2010	1245	1150	1349	1246	1172	1284	1231	119
486	West Bengal	2011	1350	1179	1314	1148	1220	1241	1185	107
487	West Bengal	2012	1346	1383	1357	1270	1352	1434	1349	120
488	West Bengal	2013	1564	1382	1474	1392	1629	1391	1315	120
489	West Bengal	2014	1516	1398	1473	1385	1527	1439	1416	135

Data By Time

In [55]:
```
databytime['STATE/UT'].unique()
```

Out[55]:
```
array(['A & N Islands', 'Andhra Pradesh', 'Arunachal Pradesh', 'Assa
m',
       'Bihar', 'Chandigarh', 'Chhattisgarh', 'D & N Haveli',
       'Daman & Diu', 'Delhi (Ut)', 'Goa', 'Gujarat', 'Haryana',
       'Himachal Pradesh', 'Jammu & Kashmir', 'Jharkhand', 'Karnatak
a',
       'Kerala', 'Lakshadweep', 'Madhya Pradesh', 'Maharashtra',
       'Manipur', 'Meghalaya', 'Mizoram', 'Nagaland', 'Odisha',
       'Puducherry', 'Punjab', 'Rajasthan', 'Sikkim', 'Tamil Nadu',
       'Tripura', 'Uttar Pradesh', 'Uttarakhand', 'West Bengal'],
      dtype=object)
```

Data By Month

In [56]:
```
databymonth['STATE/UT'].unique()
```

Out[56]:
```
array(['A & N Islands', 'Andhra Pradesh', 'Arunachal Pradesh', 'Assa
m',
       'Bihar', 'Chandigarh', 'Chhattisgarh', 'D & N Haveli',
       'Daman & Diu', 'Delhi (Ut)', 'Goa', 'Gujarat', 'Haryana',
       'Himachal Pradesh', 'Jammu & Kashmir', 'Jharkhand', 'Karnatak
a',
       'Kerala', 'Lakshadweep', 'Madhya Pradesh', 'Maharashtra',
       'Manipur', 'Meghalaya', 'Mizoram', 'Nagaland', 'Odisha',
       'Puducherry', 'Punjab', 'Rajasthan', 'Sikkim', 'Tamil Nadu',
       'Tripura', 'Uttar Pradesh', 'Uttarakhand', 'West Bengal'],
      dtype=object)
```

This Dataset contains few duplicate names for some states and that creates inconsistency in dataset. For Example : Dadra & Nagar Haveli has data with two different name 'D & N Haveli' and 'D&N Haveli'. We need to fix such discrepancies before analysis. These are small data issues, let's fix them inside data files.

TOP 5 States where Road Accidents are highest

In [135]:
```
dm1 = databymonth.drop('YEAR', axis=1).groupby('STATE/UT').sum()
dm1.reset_index(inplace=True)
dm2 = dm1[['STATE/UT','TOTAL']]
dm_top5states = dm2.sort_values(by=['TOTAL'], ascending=False).head(5)
```

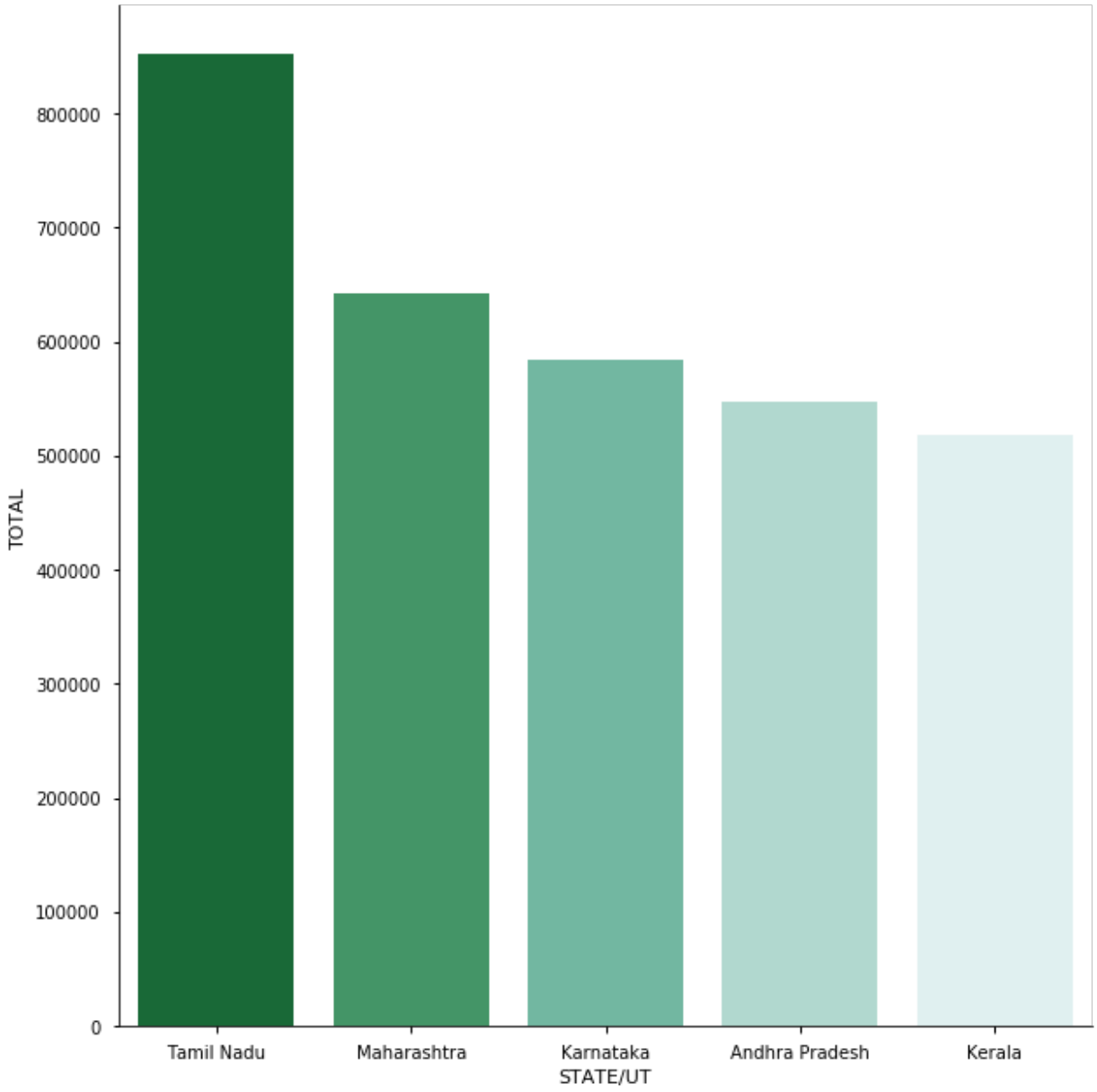
In [519]:
```
dm_top5states
```

Out[519]:

	STATE/UT	TOTAL
30	Tamil Nadu	852073
20	Maharashtra	641614
16	Karnataka	584761
1	Andhra Pradesh	546821
17	Kerala	518161

In [154]:
```
sns.factorplot(x='STATE/UT',y='TOTAL',data=dm_top5states,kind='bar', size = 9, palette='BuGn_r')
```

Out[154]: <seaborn.axisgrid.FacetGrid at 0x1a37603470>



Bottom 5 States where Road Accidents are least

In [520]:
```
dm_bottom5states = dm2.sort_values(by=['TOTAL']).head(5)
```

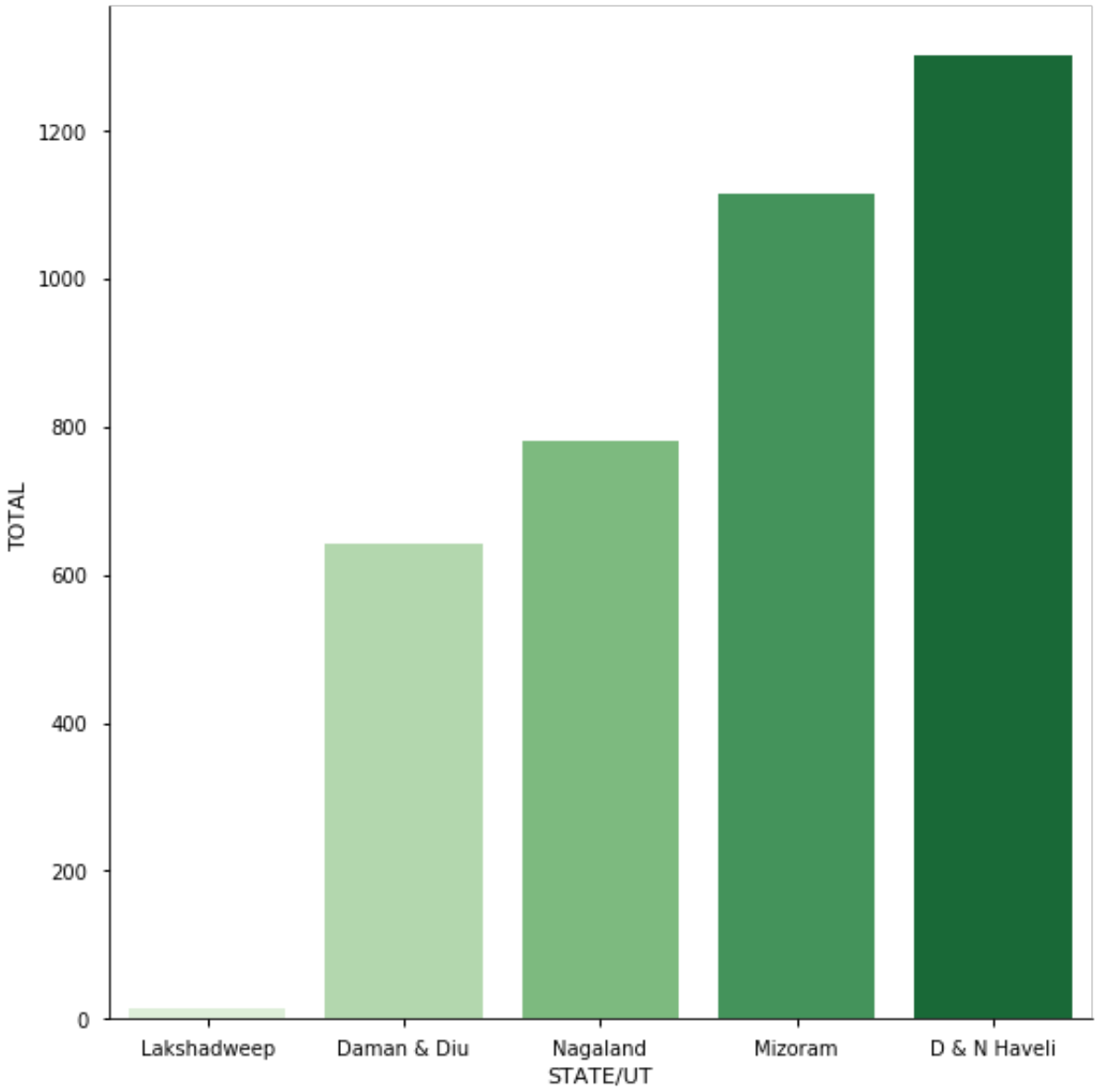
In [521]:
```
dm_bottom5states
```

Out[521]:

	STATE/UT	TOTAL
18	Lakshadweep	14
8	Daman & Diu	643
24	Nagaland	781
23	Mizoram	1116
7	D & N Haveli	1303

In [149]:
```
sns.factorplot(x='STATE/UT',y='TOTAL',data=dm_bottom5states,kind='bar'
, size = 8, palette='Greens')
```

Out[149]: <seaborn.axisgrid.FacetGrid at 0x1a34703cf8>



Data Analysis Based on Time-Window

In [523]:
```
databytime.head()
```

Out[523]:

	STATE/UT	YEAR	0-3 hrs. (Night)	3-6 hrs. (Night)	6-9 hrs (Day)	9-12 hrs (Day)	12-15 hrs (Day)	15-18 hrs (Day)	18-21 hrs (Night)	21-24 hrs (Night)	Total
0	A & N Islands	2001	2	6	29	40	39	40	18	7	181
1	A & N Islands	2002	2	6	22	41	33	33	23	8	168
2	A & N Islands	2003	2	8	31	35	28	36	25	15	180
3	A & N Islands	2004	2	5	29	42	43	43	37	14	215
4	A & N Islands	2005	0	8	27	28	38	42	50	13	206

In [590]:
```
dtime1 = databytime.drop('YEAR', axis=1).groupby('STATE/UT').sum()
dtime1.reset_index(inplace=True)
dtime1.head()
```

Out[590]:

	STATE/UT	0-3 hrs. (Night)	3-6 hrs. (Night)	6-9 hrs (Day)	9-12 hrs (Day)	12-15 hrs (Day)	15-18 hrs (Day)	18-21 hrs (Night)	21-24 hrs (Night)	Total
0	A & N Islands	32	82	388	474	607	557	552	201	2893
1	Andhra Pradesh	48306	65353	61966	71212	68608	77174	85622	68580	546821
2	Arunachal Pradesh	149	226	369	619	606	695	502	223	3389
3	Assam	2516	3280	9540	13330	11112	11293	6915	3732	61718
4	Bihar	6605	10708	13578	14666	13785	13996	11598	7712	92648

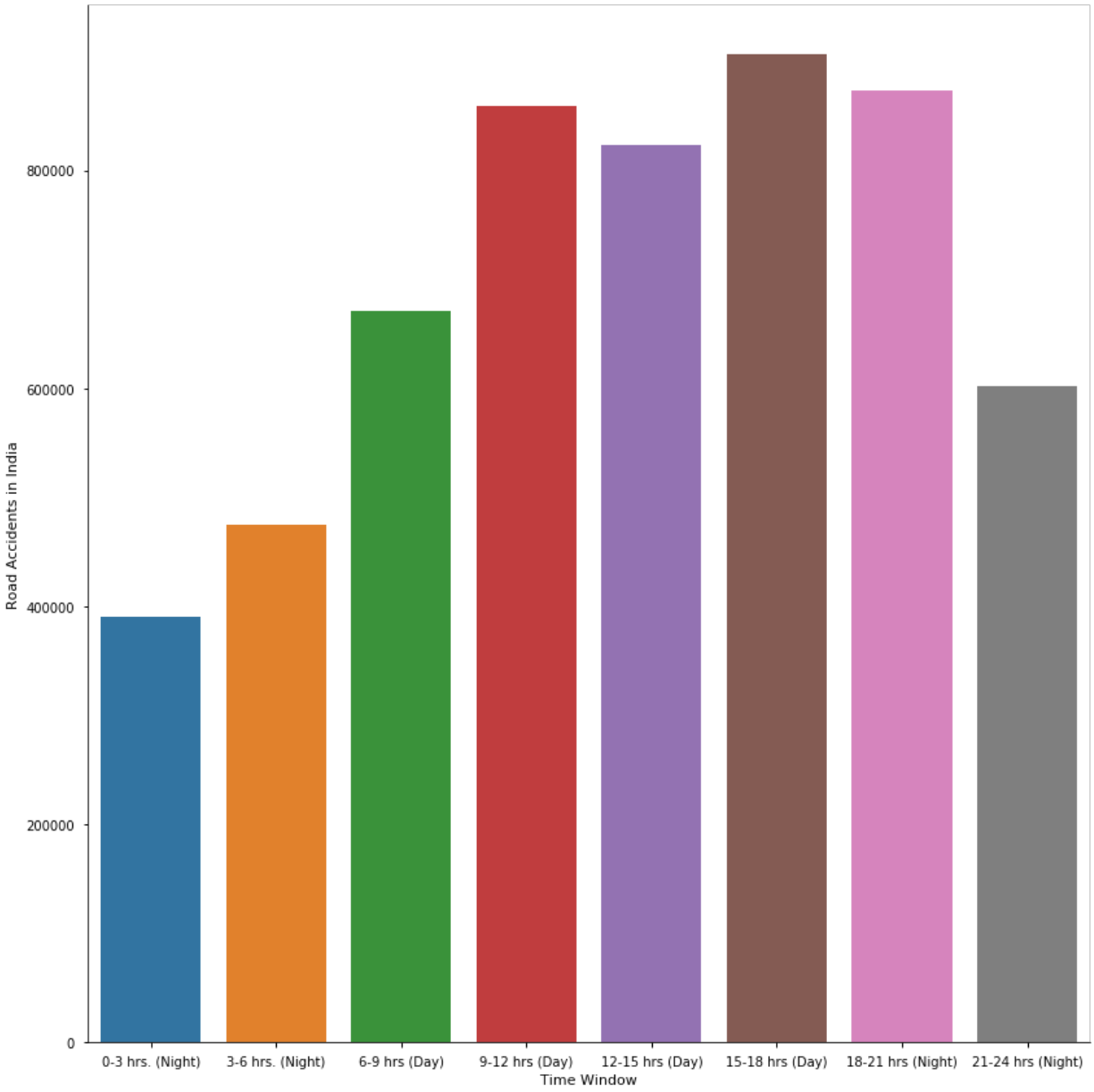
In [592]:
```
MostAccInTime=pd.DataFrame(dtime1.drop('STATE/UT', axis=1).sum())
MostAccInTime.reset_index(inplace=True)
MostAccInTime.columns = ['Time Window', 'Road Accidents in India']
MostAccInTime.drop(8, inplace=True)
MostAccInTime
```

Out[592]:

	Time Window	Road Accidents in India
0	0-3 hrs. (Night)	390197
1	3-6 hrs. (Night)	474926
2	6-9 hrs (Day)	671864
3	9-12 hrs (Day)	859444
4	12-15 hrs (Day)	824089
5	15-18 hrs (Day)	906639
6	18-21 hrs (Night)	873630
7	21-24 hrs (Night)	602117

In [593]:
```
sns.factorplot(x='Time Window',y = 'Road Accidents in India' ,data=Mos
tAccInTime,kind='bar', size = 12)
```

Out[593]: <seaborn.axisgrid.FacetGrid at 0x1a5c67fc18>



Tamil Nadu is the highest ranking state for Road Accidents

Road Accidents in Tamil Nadu by Month

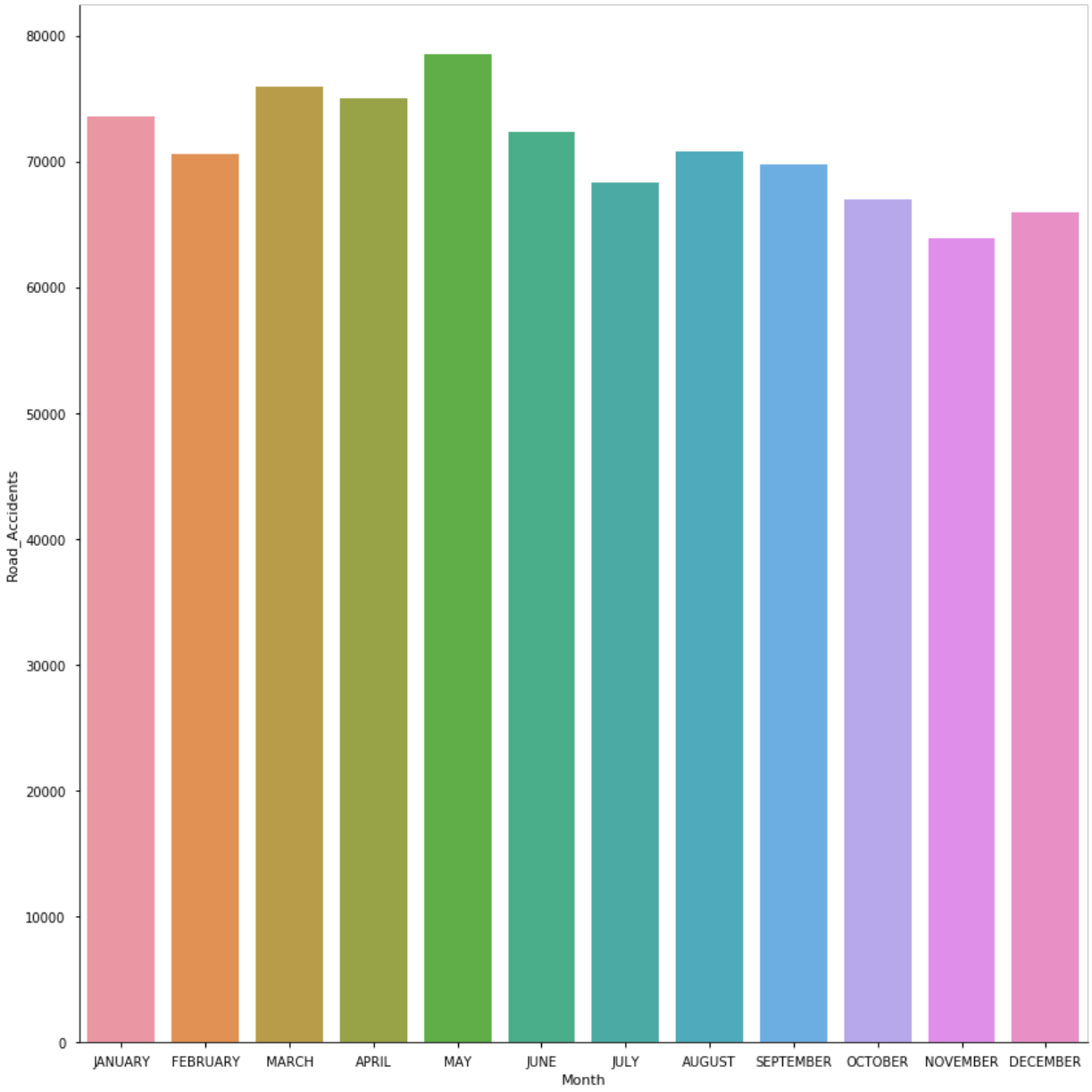
In [596]:
```
TNbymonth=dm1[dm1['STATE/UT']=='Tamil Nadu'].transpose().reset_index()
TNbymonth.columns= ['Month', 'Road_Accidents']
TNbymonth.drop(0, inplace=True)
TNbymonth.drop(13, inplace=True)
TNbymonth
```

Out[596]:

	Month	Road_Accidents
1	JANUARY	73629
2	FEBRUARY	70615
3	MARCH	75993
4	APRIL	75007
5	MAY	78514
6	JUNE	72387
7	JULY	68362
8	AUGUST	70775
9	SEPTEMBER	69796
10	OCTOBER	67053
11	NOVEMBER	63934
12	DECEMBER	66008

In [597]:
```
sns.factorplot(x='Month',y = 'Road_Accidents' ,data=TNbymonth,kind='bar', size = 12)
```

Out[597]: <seaborn.axisgrid.FacetGrid at 0x1a5ccffc18>



Road Accidents in Tamil Nadu By Year (2001 to 2014)

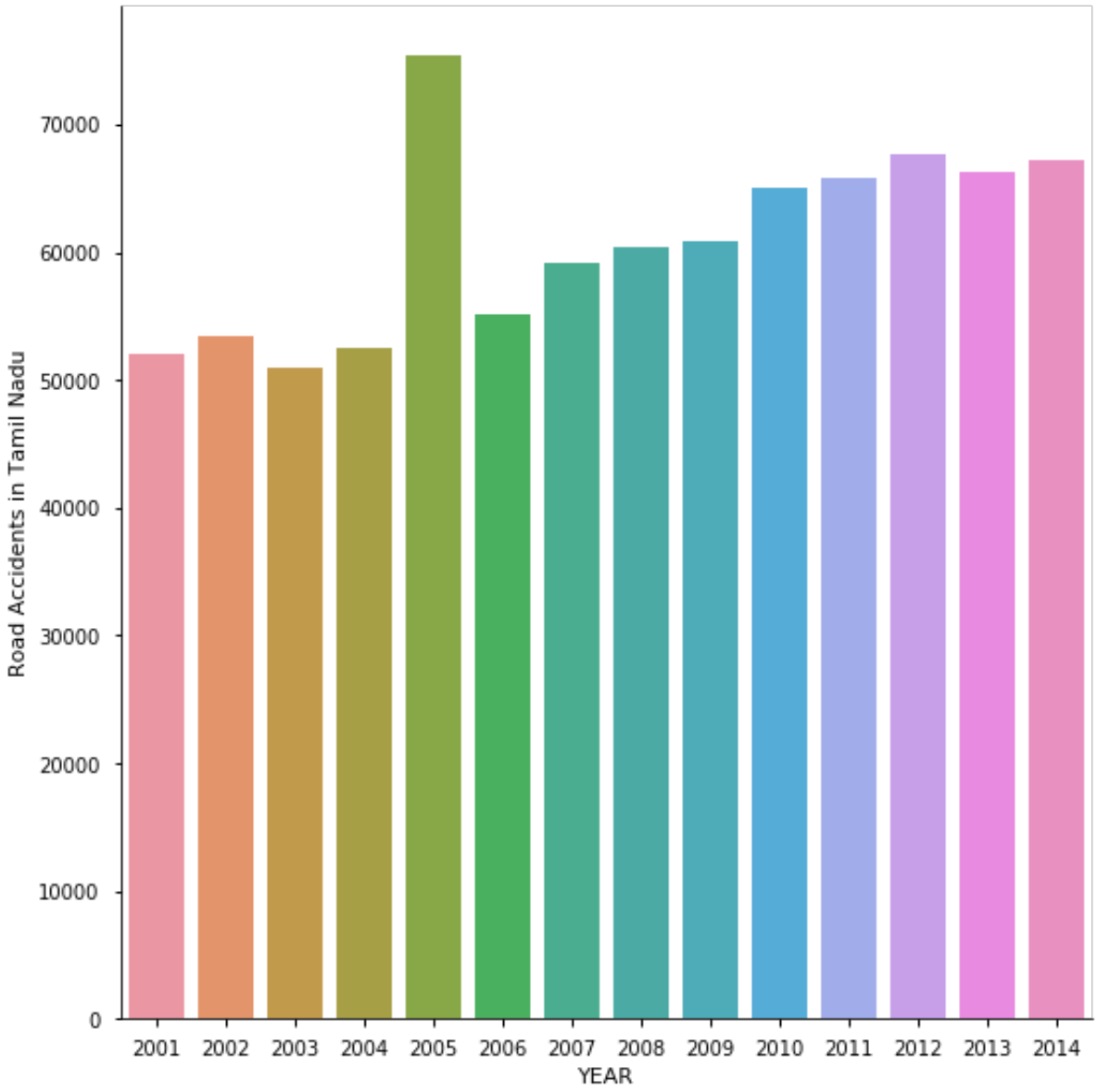
In [598]:
```
TNbyYear= dataByMonth[dataByMonth['STATE/UT']=='Tamil Nadu'][['YEAR',
'TOTAL']]
TNbyYear.columns = ['YEAR', 'Road Accidents in Tamil Nadu']
TNbyYear
```

Out[598]:

	YEAR	Road Accidents in Tamil Nadu
420	2001	51978
421	2002	53503
422	2003	51025
423	2004	52508
424	2005	75480
425	2006	55145
426	2007	59117
427	2008	60409
428	2009	60794
429	2010	64996
430	2011	65873
431	2012	67757
432	2013	66238
433	2014	67250

In [599]:
```
sns.factorplot(x='YEAR',y = 'Road Accidents in Tamil Nadu' ,data=TNbyY
ear,kind='bar', size = 8)
```

```
Out[599]: <seaborn.axisgrid.FacetGrid at 0x1a5c6d1208>
```



Road Accidents in Tamil Nadu By Time Window of Day

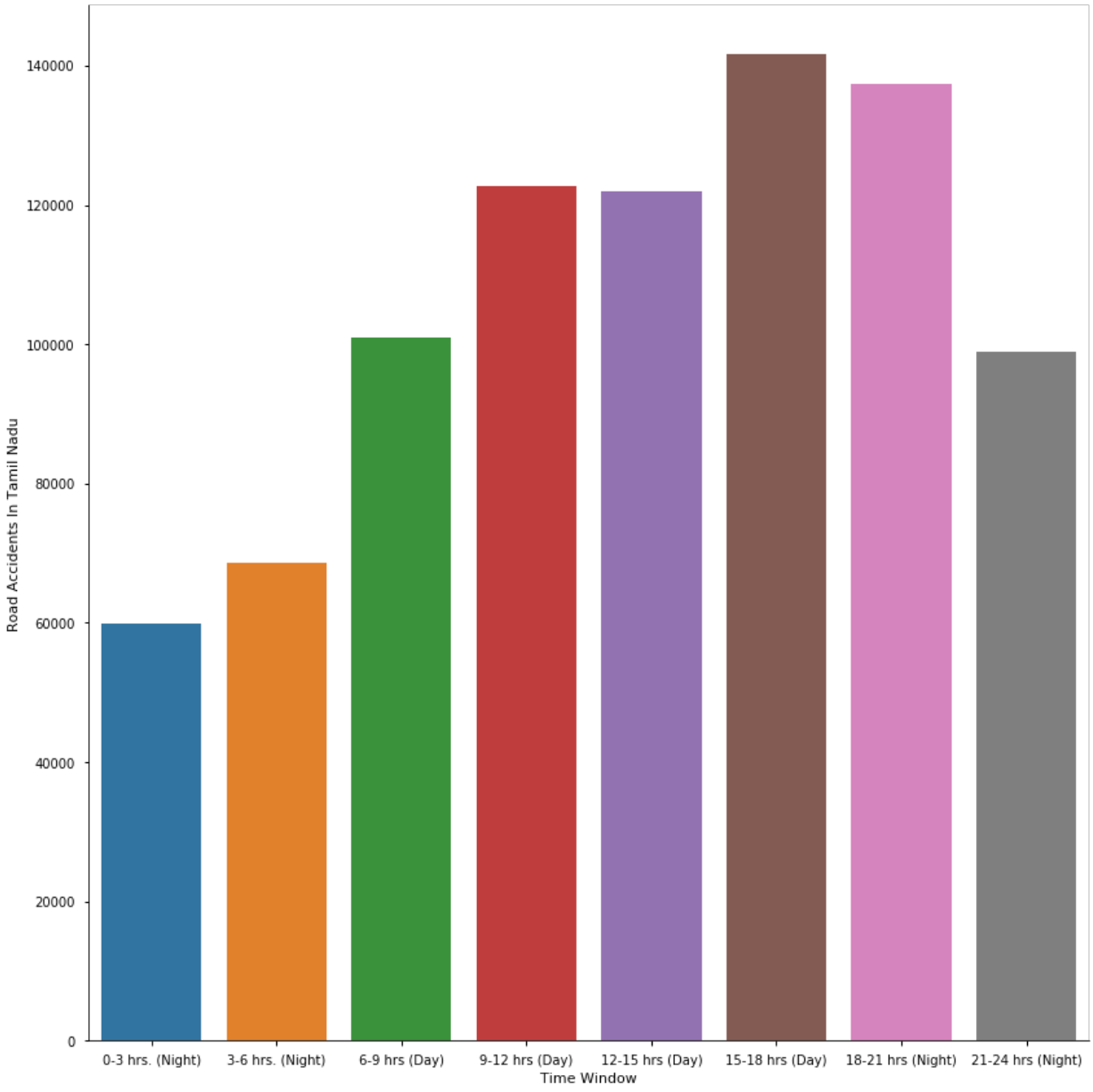
In [600]:
```
dtime1 = databytime.drop('YEAR', axis=1).groupby('STATE/UT').sum()
dtime1.reset_index(inplace=True)
TNbytime=dtime1[dtime1['STATE/UT']=='Tamil Nadu'].transpose().reset_in
dex()
TNbytime.columns= ['Time Window', 'Road Accidents In Tamil Nadu']
TNbytime.drop(0, inplace=True)
TNbytime.drop(9, inplace=True)
TNbytime
```

Out[600]:

	Time Window	Road Accidents In Tamil Nadu
1	0-3 hrs. (Night)	59955
2	3-6 hrs. (Night)	68713
3	6-9 hrs (Day)	100887
4	9-12 hrs (Day)	122724
5	12-15 hrs (Day)	122016
6	15-18 hrs (Day)	141600
7	18-21 hrs (Night)	137306
8	21-24 hrs (Night)	98872

In [601]:
```
sns.factorplot(x='Time Window',y = 'Road Accidents In Tamil Nadu' ,dat
a=TNbytime,kind='bar', size = 12)
```

Out[601]: <seaborn.axisgrid.FacetGrid at 0x1a5d6acdd8>



Road Accidents in India By Month

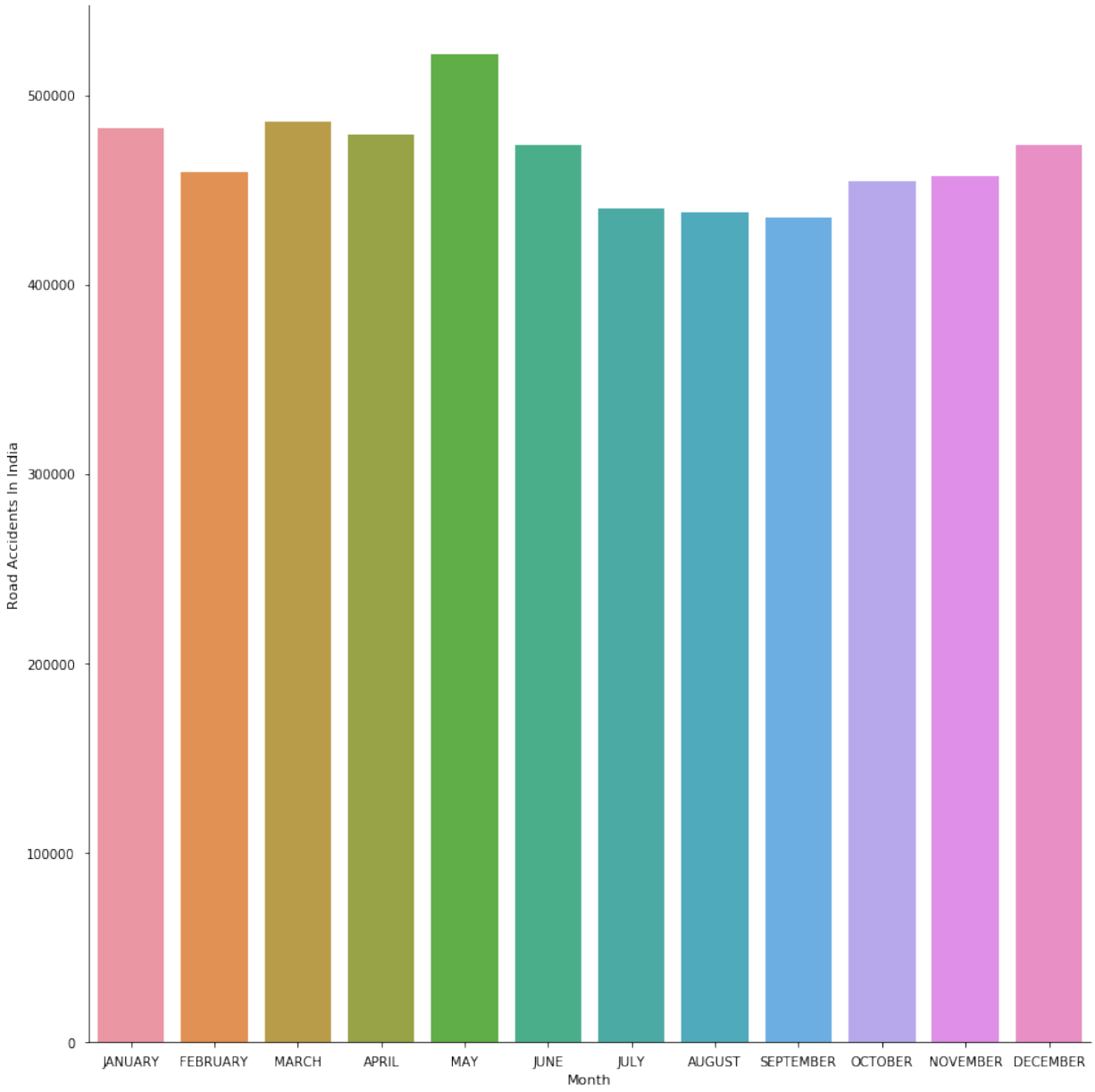
In [586]:
```
MonthData = pd.DataFrame(databymonth.drop('STATE/UT', axis=1).groupby(
'YEAR').sum().sum())
MonthData.reset_index(inplace= True)
MonthData.columns = ['Month', 'Road Accidents In India']
MonthData.drop(12, inplace=True)
MonthData
```

Out[586]:

	Month	Road Accidents In India
0	JANUARY	482719
1	FEBRUARY	459272
2	MARCH	486141
3	APRIL	479663
4	MAY	521563
5	JUNE	473574
6	JULY	440263
7	AUGUST	438351
8	SEPTEMBER	435302
9	OCTOBER	454961
10	NOVEMBER	457192
11	DECEMBER	473905

In [587]:
```
sns.factorplot(x='Month',y = 'Road Accidents In India' ,data=MonthData
,kind='bar', size = 12)
```

Out[587]:
```
<seaborn.axisgrid.FacetGrid at 0x1a5b812e80>
```



Road Accidents in India By Year

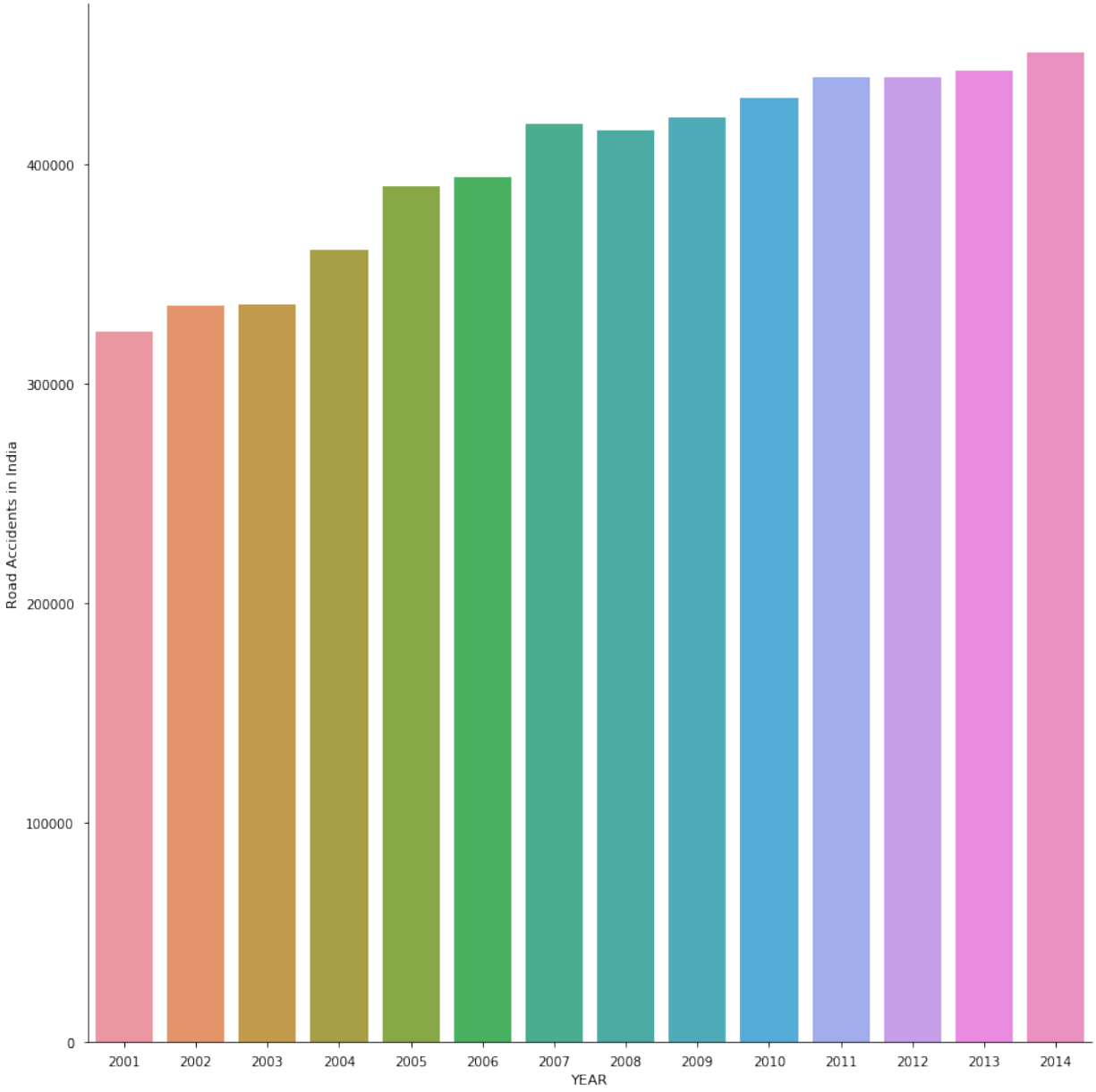
In [588]:
```
MostAccInYear = databymonth[['YEAR','TOTAL']].groupby('YEAR').sum()
MostAccInYear.reset_index(inplace=True)
MostAccInYear.columns=['YEAR', 'Road Accidents in India']
MostAccInYear
```

Out[588]:

	YEAR	Road Accidents in India
0	2001	323720
1	2002	335707
2	2003	336468
3	2004	361343
4	2005	390378
5	2006	394432
6	2007	418657
7	2008	415855
8	2009	421628
9	2010	430654
10	2011	440123
11	2012	440042
12	2013	443001
13	2014	450898

In [589]:
```
sns.factorplot(x='YEAR',y = 'Road Accidents in India' ,data=MostAccInY
ear,kind='bar', size = 12)
```

```
Out[589]: <seaborn.axisgrid.FacetGrid at 0x1a5bea80b8>
```



Heatmap to show Road Accidents based on Year for all States.

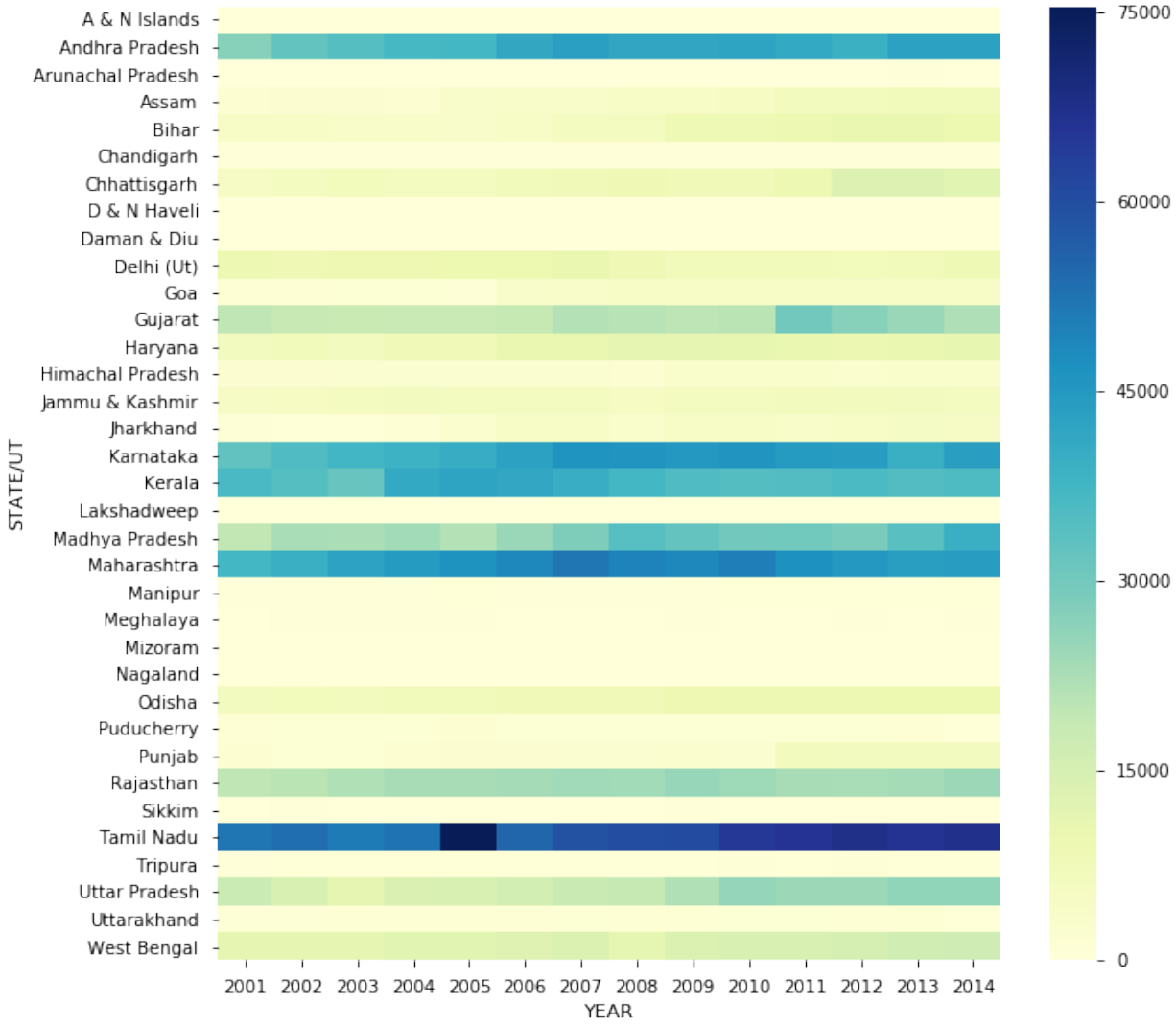
In [583]:
```
road_heatmap = databymonth.pivot_table(values='TOTAL',index='STATE/UT'
,columns='YEAR')
road_heatmap.head()
```

Out[583]:

YEAR	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	20
STATE/UT											
A & N Islands	181	168	180	215	206	155	152	191	271	285	235
Andhra Pradesh	27188	32577	34537	37078	37289	41323	43594	42106	42011	42428	410
Arunachal Pradesh	236	235	229	217	237	243	230	261	261	280	263
Assam	1806	2243	2145	2002	3656	4080	3995	4262	4585	5485	656
Bihar	4334	4372	3902	3890	3746	4382	5631	6180	8366	8441	935

In [584]:
```
plt.subplots(figsize = (10, 10))
sns.set_context("notebook", font_scale=1, rc={"lines.linewidth": 10.9}
)
sns.heatmap(road_heatmap,cmap="YlGnBu")
```

Out[584]: <matplotlib.axes._subplots.AxesSubplot at 0x1a5b8a4a58>



Heatmap to show Road Accidents based on Month for all States.

In [610]:
```
dd1 = databymonth.drop('YEAR', axis=1).groupby('STATE/UT').sum()
dd1.reset_index(inplace=True)
dd1.drop('TOTAL', axis=1, inplace=True)
dd1.columns = ['STATE', 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
dd1.head()
```

Out[610]:

	STATE	1	2	3	4	5	6	7	8	9	10	
0	A & N Islands	264	246	248	256	211	238	217	234	238	267	20
1	Andhra Pradesh	46452	44657	47732	46893	51442	46570	42374	43947	41520	43697	44
2	Arunachal Pradesh	306	298	295	253	298	263	238	247	284	290	30
3	Assam	5318	5207	5623	5224	5122	4961	4845	4890	4837	5319	52
4	Bihar	7075	7618	8538	7760	9984	9305	7487	6714	6335	6929	76

In [611]:
```
road_heatmap_2 = dd1.pivot_table(index='STATE')
road_heatmap_2.head(5)
```

Out[611]:

	1	2	3	4	5	6	7	8	9	10	
STATE											
A & N Islands	264	246	248	256	211	238	217	234	238	267	205
Andhra Pradesh	46452	44657	47732	46893	51442	46570	42374	43947	41520	43697	440
Arunachal Pradesh	306	298	295	253	298	263	238	247	284	290	301
Assam	5318	5207	5623	5224	5122	4961	4845	4890	4837	5319	522
Bihar	7075	7618	8538	7760	9984	9305	7487	6714	6335	6929	760

In [612]:
```
road_heatmap_2.columns = ['JANUARY','FEBRUARY','MARCH','APRIL','MAY','
JUNE',
                          'JULY','AUGUST','SEPTEMBER','OCTOBER','NOVEM
BER','DECEMBER']
road_heatmap_2
```

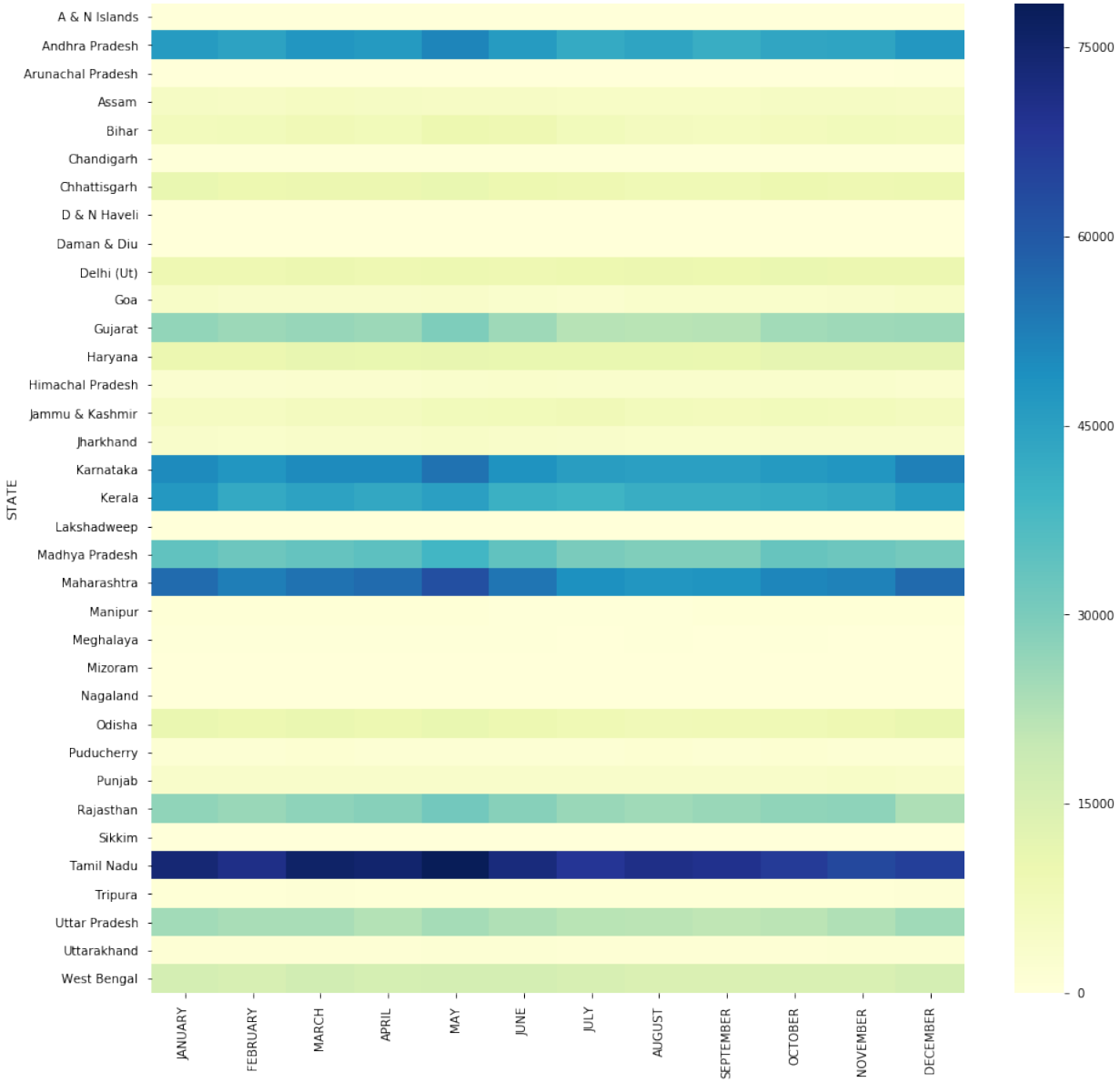
Out[612]:

	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST
STATE								
A & N Islands	264	246	248	256	211	238	217	234
Andhra Pradesh	46452	44657	47732	46893	51442	46570	42374	43947
Arunachal Pradesh	306	298	295	253	298	263	238	247
Assam	5318	5207	5623	5224	5122	4961	4845	4890
Bihar	7075	7618	8538	7760	9984	9305	7487	6714
Chandigarh	521	498	549	525	547	484	491	532
Chhattisgarh	10826	10040	10157	10337	10992	9847	9269	8621
D & N Haveli	123	122	120	96	112	110	92	97
Daman & Diu	72	48	54	62	47	48	56	40
Delhi (Ut)	9282	9355	10289	9470	9806	9438	9982	10295
Goa	4365	3813	3801	3707	4179	3290	3013	3377
Gujarat	26816	25489	26091	25632	29517	25311	21979	21732
Haryana	10198	10265	11229	10878	11244	10799	10892	10792
Himachal Pradesh	2584	2456	2880	2945	3309	3432	3350	3271
Jammu & Kashmir	5623	5586	6513	6601	7434	7830	8026	7198
Jharkhand	3692	3646	4050	4098	4552	4145	3712	3499
Karnataka	50031	47705	50224	50130	55013	48750	45768	45306
Kerala	47056	42812	44549	43224	45178	40970	39633	41431
Lakshadweep	4	1	0	1	0	0	1	0
Madhya Pradesh	34303	32414	33594	34858	38998	34211	30226	29240
Maharashtra	56243	52476	55072	56174	62394	54431	48944	47629
Manipur	740	693	727	688	631	589	595	571
Meghalaya	341	314	373	324	347	314	303	339
Mizoram	115	86	123	88	97	70	102	75

Nagaland	71	82	72	69	56	75	71	51
Odisha	10517	9591	10439	9702	10829	9755	8987	8279
Puducherry	1656	1683	1911	1740	1815	1795	1833	1858
Punjab	3753	3843	3855	3865	3791	3824	3783	3683
Rajasthan	27117	25947	27498	28249	31539	28471	26012	25041
Sikkim	236	203	256	253	211	240	172	210
Tamil Nadu	73629	70615	75993	75007	78514	72387	68362	70775
Tripura	902	854	931	831	893	838	845	842
Uttar Pradesh	24819	23644	23888	22178	24403	22808	21504	21282
Uttarakhand	1410	1428	1503	1522	1696	1686	1533	1369
West Bengal	16259	15537	16964	16023	16362	16289	15566	14884

In [613]:
```
plt.subplots(figsize = (15, 15))
sns.set_context("notebook", font_scale=1, rc={"lines.linewidth": 10.9}
)
sns.heatmap(road_heatmap_2,cmap="YlGnBu")
```

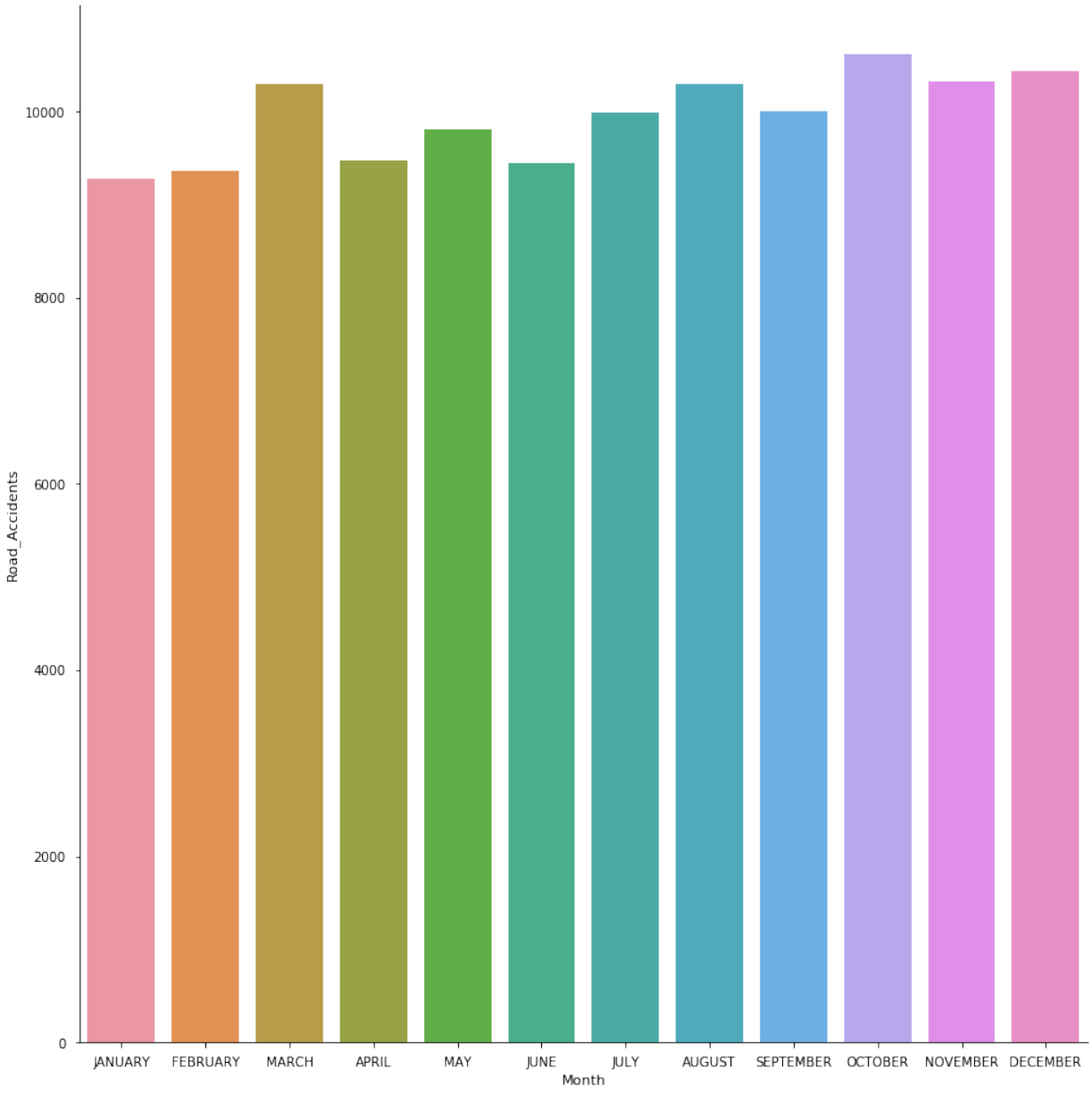
Out[613]: <matplotlib.axes._subplots.AxesSubplot at 0x1a5df8a4a8>



Road Accidents in Delhi

In [629]:
```
UTbymonth=dm1[dm1['STATE/UT']=='Delhi (Ut)'].transpose().reset_index()
UTbymonth.columns= ['Month', 'Road_Accidents']
UTbymonth.drop(0, inplace=True)
UTbymonth.drop(13, inplace=True)
sns.factorplot(x='Month',y = 'Road_Accidents' ,data=UTbymonth,kind='bar', size = 12)
```

Out[629]: <seaborn.axisgrid.FacetGrid at 0x1a64347588>



Road Accidents in Lakshadweep

In [630]:
```
UTbymonth=dm1[dm1['STATE/UT']=='Lakshadweep'].transpose().reset_index(
)
UTbymonth.columns= ['Month', 'Road_Accidents']
UTbymonth.drop(0, inplace=True)
UTbymonth.drop(13, inplace=True)
#ut=databytime[databytime['STATE/UT']=='Delhi (Ut)'].drop('YEAR',axis=
1)
#ut.groupby('STATE/UT').sum()
sns.factorplot(x='Month',y = 'Road_Accidents' ,data=UTbymonth,kind='ba
r', size = 12)
```

Out[630]: <seaborn.axisgrid.FacetGrid at 0x1a66bde470>